The Annals of Statistics
2003, Vol. 31, No. 5, 1636–1668
© Institute of Mathematical Statistics, 2003

DIMENSION REDUCTION FOR THE CONDITIONAL MEAN IN REGRESSIONS WITH CATEGORICAL PREDICTORS

BY BING LI,[1] R. DENNIS COOK[2] AND FRANCESCA CHIAROMONTE

Pennsylvania State University, University of Minnesota and Pennsylvania State University

Consider the regression of a response Y on a vector of quantitative predictors $\mathbf{X}$ and a categorical predictor W. In this article we describe a first method for reducing the dimension of $\mathbf{X}$ without loss of information on the conditional mean $\mathrm{E}(Y|\mathbf{X}, W)$ and without requiring a prespecified parametric model. The method, which allows for, but does not require, parametric versions of the subpopulation mean functions $\mathrm{E}(Y|\mathbf{X}, W = w)$, includes a procedure for inference about the dimension of $\mathbf{X}$ after reduction. This work integrates previous studies on dimension reduction for the conditional mean $\mathrm{E}(Y|\mathbf{X})$ in the absence of categorical predictors and dimension reduction for the full conditional distribution of $Y|(\mathbf{X}, W)$. The methodology we describe may be particularly useful for constructing low-dimensional summary plots to aid in model-building at the outset of an analysis. Our proposals provide an often parsimonious alternative to the standard technique of modeling with interaction terms to adapt a mean function for different subpopulations determined by the levels of W. Examples illustrating this and other aspects of the development are presented.

1. Introduction. A common paradigm for studying the regression of a univariate response Y on a vector $\mathbf{X} \in \mathbb{R}^p$ of quantitative continuous or many-valued predictors hinges on describing the conditional distribution of $Y|\mathbf{X}$ with a parsimonious parametric model. Depending on available data and study-specific goals, modeling may be restricted to the conditional mean $\mathrm{E}(Y|\mathbf{X})$ and perhaps the conditional variance $\mathrm{Var}(Y|\mathbf{X})$, leaving other aspects of $Y|\mathbf{X}$ unspecified or to be filled in by plausible assumptions.

When a parametric model for $Y|\mathbf{X}$ is not available ex ante and the dimension of $\mathbf{X}$ is too large for direct visualization of the data, the theory of *sufficient dimension reduction* may provide an effective starting point for the regression. This theory provides a framework for replacing $\mathbf{X}$ with a lower dimensional linearly transformed version $\mathbf{A}'\mathbf{X}$ without loss of information on targeted characteristics of the conditional distribution of $Y|\mathbf{X}$ and without requiring a prespecified parametric model. Model building can then be limited to the reduced predictors $\mathbf{A}'\mathbf{X}$ expressed

Received August 2001; revised June 2002.

[1]Supported in part by NSF Grant DMS-02-04662.

[2]Supported in part by NSF Grant DMS-01-03983.

AMS 2000 subject classifications. Primary 62G08; secondary 62G09, 62H05.

Key words and phrases. Analysis of covariance, central subspace, graphics, OLS, SIR, PHD, SAVE.

as linear combinations of the original ones. The drop in dimension is often substantial in practice, even when starting with a high-dimensional $\mathbf{X}$. Reduced predictors with dimension at most three suffice in many applications, and allow a fully informative and direct visualization of the original regression through a plot of Y versus $\mathbf{A}'\mathbf{X}$.

Let $W \in \{1, \ldots, c\}$ denote a categorical predictor that partitions the population into c subpopulations. The variable W could represent a qualitative predictor like species, or a combination of qualitative predictors like species and location, or it could be a categorical version of a continuous predictor. In this article we introduce a dimension reduction method for $\mathbf{X}$ in the conditional mean $\mathrm{E}(Y|\mathbf{X}, W)$. We assume throughout that the data are i.i.d. observations on $(\mathbf{X}, Y, W)$, which has a joint distribution. Developments and proofs can be modified straightforwardly to accommodate the case where W is nonrandom and $(\mathbf{X}, Y)|W$ has a joint distribution for each level of W. The methodology to be described applies without modification when W is nonrandom.

1.1. *Dimension reduction for $Y|\mathbf{X}$.* A dimension reduction subspace for $Y|\mathbf{X}$ is any subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}, \tag{1}$$

where $\mathbf{P}_{\cdot}$ stands for a projection operator and $\perp\!\!\!\perp$ indicates independence. The statement is thus that Y is independent of $\mathbf{X}$ given $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Under mild conditions the intersection of all dimension reduction subspaces also satisfies (1), and in these cases it is called the *central subspace* (CS) of the regression and indicated with $\mathcal{S}_{Y|\mathbf{X}}$. The CS, which represents the minimal subspace that preserves the original information on $Y|\mathbf{X}$, is the main object of interest for reducing the dimension of $\mathbf{X}$ without loss of information on $Y|\mathbf{X}$. It is unique when it exists and thus constitutes a well defined object of inference.

A summary plot of Y versus $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ is called the *central view* of the regression, with the understanding that it may be directly visualizable only when $\dim(\mathcal{S}_{Y|\mathbf{X}})$ is small. In practice, we can display an estimated central view in terms of any basis $(\mathbf{s}_1, \ldots, \mathbf{s}_d)$ for an estimate of $\mathcal{S}_{Y|\mathbf{X}}$ by plotting Y versus $\mathbf{s}_1'\mathbf{X}, \ldots, \mathbf{s}_d'\mathbf{X}$. Uncorrelated views with the sample $\mathrm{Var}(\mathbf{s}_1'\mathbf{X}, \ldots, \mathbf{s}_d'\mathbf{X}) = I$ often provide the best visual resolution.

There are several methods for estimating $\mathcal{S}_{Y|\mathbf{X}}$ or portions thereof under restrictions on the marginal distribution of $\mathbf{X}$, including *ordinary least squares* [OLS; Li and Duan (1989)], *sliced inverse regression* [SIR; Li (1991)], *sliced average variance estimation* [SAVE; Cook and Weisberg (1991)], *principal Hessian directions* [PHD; Li (1992); see also Cook (1998b)] and *parametric inverse regression* [PIR; Bura and Cook (2001)].

Recent advances have expanded the scope of sufficient dimension reduction in two fundamental directions.

1.2. *Dimension reduction for* $\mathrm{E}(Y|\mathbf{X})$. Cook and Li (2002) investigated dimension reduction for $\mathrm{E}(Y|\mathbf{X})$. A dimension reduction subspace for the conditional mean is any subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\mathbf{P}_{\mathcal{S}}\mathbf{X}. \tag{2}$$

When the intersection of all subspaces satisfying (2) does itself satisfy the condition, it is called the *central mean subspace* (CMS) of the regression and is indicated with $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$. It is straightforward to show that $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ with equality in the case of *location regressions* where $Y \perp\!\!\!\perp \mathbf{X}|\mathrm{E}(Y|\mathbf{X})$. Because of this inclusion, any method of estimating directions within the CMS will estimate directions within the CS. Conversely, Cook and Li proved that, among existing methods to estimate directions within the CS, PHD and OLS always estimate directions within $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$, while SIR and SAVE can estimate directions in $\mathcal{S}_{Y|\mathbf{X}}$ but not in $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$. After $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$ has been estimated, features of the mean function can be studied in a summary plot of Y versus $\mathbf{s}_1'\mathbf{X}, \ldots, \mathbf{s}_d'\mathbf{X}$, where $\{\mathbf{s}_1, \ldots, \mathbf{s}_d\}$ is a basis for the estimate of $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.

1.3. *Partial dimension reduction for* $Y|(\mathbf{X}, W)$. The discussion of dimension reduction in Sections 1.1 and 1.2 was limited to regressions with quantitative predictors $\mathbf{X}$ because it is in such settings that linear dimension reduction may be particularly relevant. Straightforward application to regressions that include a categorical predictor W may be inappropriate because then the relevance of linear combinations involving W can be elusive.

Chiaromonte, Cook and Li (2002) investigated the reduction of $\mathbf{X}$ in regressions that include a categorical predictor $W \in \{1, \ldots, c\}$. A partial dimension reduction subspace is any subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W). \tag{3}$$

If the intersection of all such partial subspaces itself satisfies (3), it is called the *central partial subspace* (CPS) for the regression of Y on $(\mathbf{X}, W)$ and is indicated with $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$. For partial dimension reduction, the summary view is a plot of Y versus $\mathbf{P}_{\mathcal{S}^{(W)}_{Y|\mathbf{X}}}\mathbf{X}$ with points marked to indicate the W subpopulation.

Let $(\mathbf{X}_w, Y_w)$ indicate a pair distributed like $(\mathbf{X}, Y)|(W = w)$ and let $\mathcal{S}_{Y_w|\mathbf{X}_w}$ denote the central space for the regression of Y_w on $\mathbf{X}_w$. Also, let $\oplus$ indicate the direct sum between two subspaces ($V_1 \oplus V_2 = \{v_1 + v_2; v_1 \in V_1, v_2 \in V_2\}$). Chiaromonte, Cook and Li (2002) proved that

$$\mathcal{S}^{(W)}_{Y|\mathbf{X}} = \bigoplus_{w=1}^{c} \mathcal{S}_{Y_w|\mathbf{X}_w}. \tag{4}$$

Although the spaces $\mathcal{S}_{Y_w|\mathbf{X}_w}$, $w = 1, \ldots, c$, can overlap in any fashion, the CPS always coincides with their direct sum. This suggests that partial dimension

reduction be performed by combining dimension reduction within subpopulations. In particular, Chiaromonte, Cook and Li (2002) adapted SIR for estimation of directions in $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$.

1.4. *Overview.* In this article, we merge the two lines of inquiry described in Sections 1.2 and 1.3, studying dimension reduction for conditional means involving a categorical predictor. We formalize this notion in Section 2 and map its connections to the ideas reviewed in Sections 1.1–1.3. We study the corresponding estimation problem in Section 3. Section 4 contains the large sample results that allow us to estimate the dimension of the relevant subspace under various assumptions. Section 5 describes an implementation of the proposed methodology. Section 6 introduces a more efficient estimator by pooling observations from different categories, and Section 7 contains two applications that illustrate its potential benefits. We give some final remarks in Section 8. Proofs for most propositions are provided in a technical appendix. It is important to note that, although our categorical variable W represents a single category throughout the article, it covers the multiple category case because a multiple categorical predictor can be represented by a single categorical predictor with multiple levels of classes.

2. Partial dimension reduction for $\mathrm{E}(Y|\mathbf{X}, W)$. In this section we begin the development of new methods for reducing the dimension of $\mathbf{X}$ in the mean function $\mathrm{E}(Y|\mathbf{X}, W)$. A partial dimension reduction subspace for this mean is any subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X}, W)|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W). \tag{5}$$

Accordingly, $(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W)$ contains all the information that the predictor $(\mathbf{X}, W)$ has to furnish on the mean function $\mathrm{E}(Y|\mathbf{X}, W)$. This conditional independence statement can be reexpressed:

PROPOSITION 2.1. *Condition* (5) *is equivalent to either of the following*:

1. $\mathrm{Cov}(Y, \mathrm{E}(Y|\mathbf{X}, W)|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W) = 0$.
2. $\mathrm{E}(Y|\mathbf{X}, W) = \mathrm{E}(Y|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W)$.

Proposition 2.1.1 allows us to understand a partial dimension reduction subspace for the mean in terms of conditional subpopulation correlations between Y and $\mathrm{E}(Y|\mathbf{X}, W)$: The projection $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ is sufficient for the mean function if and only if, within each subpopulation determined by W, Y and $\mathrm{E}(Y|\mathbf{X}, W)$ are uncorrelated given $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Proposition 2.1.2 confirms the intuition that $\mathrm{E}(Y|\mathbf{X}, W)$ depends on $\mathbf{X}$ only through $\mathbf{P}_{\mathcal{S}}\mathbf{X}$.

For convenience, we will often use the abbreviation

$$\mathrm{E}\big(f(\mathbf{X}, Y)|g(\mathbf{X}), W = w\big) \equiv \mathrm{E}\big(f(\mathbf{X}_w, Y_w)|g(\mathbf{X}_w)\big),$$

where f and g are arbitrary functions. For example, $\mathrm{E}(Y|\boldsymbol{\alpha}'\mathbf{X}, W = w)$ will often be written as $\mathrm{E}(Y_w|\boldsymbol{\alpha}'\mathbf{X}_w)$.

2.1. *Central partial mean subspace.* Assuming that the intersection of all subspaces satisfying (5) does itself satisfy (5), we obtain a unique object of inference. We call this intersection the *central partial mean subspace* (CPMS), and indicate it with $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$. A summary plot of Y versus $\mathbf{P}_{\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}}\mathbf{X}$ with points marked to indicate the W subpopulations is called the *central partial mean view*.

Like the three types of central subspaces reviewed in Sections 1.1–1.3, some constraints on the regression are required to insure that the CPMS exists. The known constraints used to guarantee the existence of the central subspace also guarantee the existence of $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$. Since these conditions accommodate a very broad range of practical applications, we will always assume the CPMS to exist. For background on the existence issue for central subspaces, see Cook [(1996) and (1998a), Chapter 6] and Chiaromonte and Cook (2002).

It is straightforward to prove that the CPMS is contained in the central partial space:

$$\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})} \subseteq \mathcal{S}^{(W)}_{Y|\mathbf{X}}.$$

Because of this inclusion, any method estimating directions within the CPMS will also estimate directions within the CPS.

2.2. *Location regressions.* The CPMS is the same as the CPS $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})} = \mathcal{S}^{(W)}_{Y|\mathbf{X}}$ within the class of *location regressions* defined by the relation $Y \perp\!\!\!\perp (\mathbf{X}, W) | \mathrm{E}(Y|\mathbf{X}, W)$. This class covers many models that are useful in practice. For instance, the additive error models with $\varepsilon \perp\!\!\!\perp (\mathbf{X}, W)$, $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = 1$ are all location regressions,

$$Y_w = \mu_w + \boldsymbol{\eta}'\mathbf{X}_w + \sigma\varepsilon, \tag{6}$$

$$= \mu_w + \boldsymbol{\eta}_w'\mathbf{X}_w + \sigma_w\varepsilon, \tag{7}$$

$$= g_w(\boldsymbol{\eta}'\mathbf{X}_w) + \sigma_w(\boldsymbol{\eta}'\mathbf{X}_w)\varepsilon, \tag{8}$$

$$= g_w(\boldsymbol{\eta}_w'\mathbf{X}_w) + \sigma_w(\boldsymbol{\eta}_w'\mathbf{X}_w)\varepsilon, \tag{9}$$

$$= g_w(\mathbf{H}_w'\mathbf{X}_w) + \sigma\varepsilon, \tag{10}$$

for $w = 1, \ldots, c$. Model (6) is the standard homoscedastic analysis of covariance model with $\boldsymbol{\eta} \in \mathbb{R}^p$. The subpopulation mean functions are all linear in $\mathbf{X}$ and can differ only in their intercepts. Model (7) stipulates a linear regression in each subpopulation, but allows the coefficient vectors $\boldsymbol{\eta}_w \in \mathbb{R}^p$ and the standard deviations σ_w to differ. In model (8), as in model (6), the subpopulation mean functions depend on $\mathbf{X}$ only through the single linear combination $\boldsymbol{\eta}'\mathbf{X}$. Otherwise these mean functions can differ rather arbitrarily as represented by the functions g_w, which may be known or unknown. The variance function in (6) is constant, but a subpopulation variance function in (8) can depend on the linear

combination of $\mathbf{X}$ that drives the corresponding mean function. In both (6) and (8), $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})} = \mathrm{Span}(\boldsymbol{\eta})$.

Models (7) and (9) are structurally equivalent in the dimension reduction context since in both cases $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})} = \mathrm{Span}(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_c)$. Although standard linear regression models are assumed to hold within subpopulations in (7), dimension reduction across subpopulations is still a relevant issue. In this way the proposals discussed herein provide an alternative to the standard technique of modeling with interaction terms to adapt a mean function for different subpopulations.

In model (10), $\mathbf{H}_w$ represents a $p \times q_w$ matrix, so the mean function for subpopulation w depends on q_w linear combinations of the predictors, and $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})} = \mathrm{Span}(\mathbf{H}_1, \ldots, \mathbf{H}_c)$. The CPMS is not restricted to describing additive error regressions as illustrated in (6)–(10). In logistic regression, for example, we might entertain a model of the form

$$\mathrm{logit}(\mathbf{X}_w) = g_w(\boldsymbol{\eta}_w' \mathbf{X}_w), \tag{11}$$

which is also a location regression.

While location regressions are important to recognize when restricting attention to the mean function $\mathrm{E}(Y|\mathbf{X}, W)$, neither the definition of the CPMS nor methods for estimating it are restricted to this class.

2.3. *Characterizing* $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$. The CPMS represents the minimal subspace of $\mathbb{R}^p$ that preserves $\mathrm{E}(Y|\mathbf{X}, W)$. Although W is not subject to reduction, the subpopulation structure it induces affects location information, and thus shapes the conditional independence relationship (5) through which the reduction of $\mathbf{X}$ is performed. The CPMS $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$ need not coincide with the marginal CMS $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$ nor with the CMS within any subpopulation. However, these various spaces are related in a fundamental way. Let $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ denote the CMS within subpopulation w, $w = 1, \ldots, c$. As discussed in Section 2.2 for the location regression (9), $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} = \mathrm{Span}(\boldsymbol{\eta}_w)$ and $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})} = \mathrm{Span}(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_c)$. This relationship, which is a CPMS equivalent of (4), holds universally:

PROPOSITION 2.2. $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})} = \bigoplus_{w=1}^{c} \mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$.

The subpopulation central mean subspaces $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$, $w = 1, \ldots, c$, which can in principle overlap in any fashion, always add up to the CPMS. Proposition 2.2 is important because it suggests a way to develop methodology to estimate the CPMS. If we have a method to estimate the subpopulation CMS $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$, then these estimates can be combined following Proposition 2.2 to form an estimate of the CPMS.

As pointed out by Cook and Li (2002), there are two established estimation methods for targeting a cental mean subspace: OLS and PHD. In the next section

we use OLS to estimate $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ and then following Proposition 2.2 we combine these estimates into an estimate of $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$. There are several reasons for choosing OLS as a first method to estimate a CPMS. First, to work well, OLS requires fewer restrictions on $\mathbf{X}$ than PHD. Second, PHD is not very effective at finding linear trends and it tends to work best when applied to residuals [Cook (1998b)]. This means that even if PHD were extended to estimating part of a CPMS, OLS may still be needed to deal effectively with linear trends. In addition, OLS has certain desirable model robustness properties, as described in the next section.

3. Using OLS. Let $\boldsymbol{\sigma}_w = \mathrm{Cov}(Y_w, \mathbf{X}_w) \in \mathbb{R}^p$ and let $\boldsymbol{\beta}_w = \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\sigma}_w$ denote the subpopulation OLS vectors, $w = 1, \ldots, c$. We assume throughout the rest of this article that each subpopulation mean function depends on at most a single linear combination $\boldsymbol{\eta}_w'\mathbf{X}_w$ of the predictors,

$$\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} = \mathrm{Span}(\boldsymbol{\eta}_w), \qquad w = 1, \ldots, c, \tag{12}$$

where $\boldsymbol{\eta}_w \in \mathbb{R}^p$. Under this assumption, the conditional expectation $\mathrm{E}(Y_w|\mathbf{X}_w)$ reduces to $\mathrm{E}(Y_w|\boldsymbol{\eta}_w'\mathbf{X}_w)$, where $\boldsymbol{\eta}_w'\mathbf{X}_w$ is a scalar, and we will write $\mathrm{E}(Y_w|\boldsymbol{\eta}_w'\mathbf{X}_w = t)$ as $\mu_w(t)$. For models (6) and (7), $\mu_w(t) = \mu_w + t$; for models (8) and (9), $\mu_w(t) = g_w(t)$; and for model (11), $\mu_w(t) = (\mathrm{expit} \circ g_w)(t)$, where expit stands for the inverse of the logit function.

Letting $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_c)$, we propose to use

$$\mathrm{Span}(\hat{\mathbf{B}}) = \mathrm{Span}(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_c)$$

to construct an estimate of $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$, where $\hat{\boldsymbol{\beta}}_w$ is the usual sample OLS coefficient vector for $\mathbf{X}_w$. For this to be reasonable we should have

$$\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} = \mathrm{Span}(\boldsymbol{\beta}_w), \qquad w = 1, \ldots, c, \tag{13}$$

since this guarantees that $\mathrm{Span}(\mathbf{B}) = \mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$ by Proposition 2.2. The rest of this section is devoted to investigating conditions under which (13) will hold.

We will establish relationship (13) under three sets of conditions. To varying degrees, these conditions constrain either the form of the regression function $\mathrm{E}(Y_w|\mathbf{X}_w)$ or the explanatory vector $\mathbf{X}_w$. First, consider the simplest case where model (7) gives a good description of the subpopulation regressions. Then $\mathrm{E}(Y_w|\mathbf{X}_w) = \mu_w + \boldsymbol{\eta}_w'\mathbf{X}_w$. We can assume without loss of generality that $\mathrm{E}(\mathbf{X}_w) = 0$: in other words, $\boldsymbol{\eta}_w'\mathrm{E}(\mathbf{X}_w)$ is absorbed by μ_w. It follows that

$$\boldsymbol{\beta}_w = \boldsymbol{\Sigma}_w^{-1}\mathrm{E}\{(Y_w - \mu_w)\mathbf{X}_w\} = \boldsymbol{\Sigma}_w^{-1}\mathrm{E}\{\mathbf{X}_w\mathrm{E}(Y_w|\mathbf{X}_w)\} = \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_w\boldsymbol{\eta}_w = \boldsymbol{\eta}_w \tag{14}$$

and hence (13) holds.

The second set of conditions substantially reduces the restriction on the shape of the regression, but requires $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\beta}_w'\mathbf{X}_w)$ to be a linear function of $\boldsymbol{\beta}_w'\mathbf{X}_w$. Consider, for each $w = 1, \ldots, c$, subpopulation objective functions of the form

$$L_w(\theta + \mathbf{h}'\mathbf{X}_w, Y_w) = -Y_w(\theta + \mathbf{h}'\mathbf{X}_w) + \phi_w(\theta + \mathbf{h}'\mathbf{X}_w)$$

for some convex functions ϕ_w, $\theta \in \mathbb{R}^1$ and $\mathbf{h} \in \mathbb{R}^p$. Let $R_w(\theta, \mathbf{h}) = \mathrm{E}\{L_w(\theta + \mathbf{h}'\mathbf{X}_w, Y_w)\}$ and

$$(\bar{\theta}_w, \bar{\mathbf{h}}_w) = \arg\min_{\theta, \mathbf{h}} R_w(\theta, \mathbf{h})$$

denote the subpopulation minimizers. As we will see, we often have enough prior or data-analytic information on the subpopulations to conclude that, for some ϕ_w,

$$\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} = \mathrm{Span}(\bar{\mathbf{h}}_w), \qquad w = 1, \ldots, c. \tag{15}$$

Furthermore, as we will see in Proposition 3.3.2, if $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\beta}_w'\mathbf{X}_w)$ is a linear function of $\boldsymbol{\beta}_w'\mathbf{X}_w$, then $\boldsymbol{\beta}_w$ and $\boldsymbol{\eta}_w$ are parallel to each other regardless of the form of ϕ_w. In this way we establish (13) through the conditions (15), as often suggested by the regression model, and a weak form of linearity of $\mathbf{X}_w$. For instance, we might conclude that each subpopulation regression is described by a logistic regression of the form (11), so (15) would hold with ϕ_w obtained from the usual logistic objective function, which is necessarily convex. Then we need to check only the linearity of $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\beta}_w'\mathbf{X}_w)$ in $\boldsymbol{\beta}_w'\mathbf{X}_w$ to justify (13).

Condition (15) is satisfied for a wide range of regression problems. For example, for generalized linear models with natural link functions, the log likelihood has the form $-L(\theta + \mathbf{h}'\mathbf{X}_w, Y_w)$ and so $R_w(\theta, \mathbf{h})$ is the expectation of the negative log likelihood. It is well known that under regularity conditions the maximizer of the expected log likelihood is the true parameter. Consequently $\bar{\mathbf{h}}_w = \boldsymbol{\eta}_w$ and (15) follows. In fact, (15) holds much more widely than the context of generalized linear models. The next two propositions assert, roughly, that if μ_w is monotone, then there is always a form of ϕ_w such that $\bar{\mathbf{h}}_w$ satisfies (15). For convenience, we state these propositions only for increasing functions, but all the results hold for decreasing functions, as discussed in the proofs.

PROPOSITION 3.1. *Suppose* (a) *condition* (12) *holds*, (b) $\mu_w(\cdot)$ *is nondecreasing and* (c) *for any* $(\theta, \mathbf{h}) \neq (0, \boldsymbol{\eta}_w)$,

$$\Pr\left(\int_{\boldsymbol{\eta}_w'\mathbf{X}_w}^{\theta+\mathbf{h}'\mathbf{X}_w} \{\mu_w(\boldsymbol{\eta}_w'\mathbf{X}_w) - \mu_w(s)\}\, ds < 0\right) > 0. \tag{16}$$

Then there is always a form of $\phi_w(\cdot)$ *such that the corresponding* $R_w(\theta, \mathbf{h})$ *has a unique minimizer* $(\bar{\theta}_w, \bar{\mathbf{h}}_w)$, *in which the vector* $\bar{\mathbf{h}}_w$ *spans the space* $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$.

Here, again, the probability of the form $\Pr(f(X_w, Y_w) \in A)$ stands for the conditional probability $\Pr(f(X, Y) \in A|W = w)$. Notice that, because $\mu_w(\cdot)$ is nondecreasing, for any t and u the integral $\int_u^t (\mu_w(u) - \mu_w(s))\, ds$ is always nonpositive. In view of this, condition (16) is a mild addition to monotonicity. The next proposition gives two sufficient conditions for (16).

PROPOSITION 3.2. *Condition* (16) *in Proposition* 3.1 *is satisfied if either of the following two conditions is satisfied*:

1. $\mathbf{X}_w$ *is a continuous random vector with* $\mathrm{Var}(\mathbf{X}_w)$ *being positive definite*; $\mu_w(\cdot)$ *is strictly increasing*.
2. $\mathbf{X}_w$ *is a continuous random vector with an open and convex support in* $\mathbb{R}^p$; $\mu_w(\cdot)$ *is continuous and nondecreasing, and is strictly increasing in an open subinterval of* $\{\boldsymbol{\eta}_w'\mathbf{x}:\mathbf{x}\in\mathcal{X}_w\}$, *where* $\mathcal{X}_w$ *is the sample space for* $\mathbf{X}_w$.

In passing from the first to the second set of conditions in this proposition, we add restrictions on the predictor vector while relaxing the requirements on μ_w.

The third set of conditions that we use to insure (13) imposes no restrictions on the shape of μ_w, but does impose a form of linearity on $\mathbf{X}_w$. In particular, if we know that $\mathrm{E}(\mathbf{X}|\boldsymbol{\eta}_w'\mathbf{X}_w)$ is linear in $\boldsymbol{\eta}_w'\mathbf{X}_w$, then, as we will see in Proposition 3.3.3, $\boldsymbol{\beta}_w$ satisfies (13) regardless of the form of the mean function and regardless of whether the regression is a location regression. However, because $\boldsymbol{\eta}_w$ is unknown, in practice we may need to require that $\mathrm{E}(\mathbf{X}|\boldsymbol{\alpha}'\mathbf{X}_w)$ be linear in $\boldsymbol{\alpha}'\mathbf{X}$ for all $\boldsymbol{\alpha}$, which is true when $\mathbf{X}$ has an elliptically contoured distribution.

The next proposition summarizes the three sets of sufficient conditions for (13). We will assume that each $\mathbf{X}_w$ is a continuous random vector with $\mathrm{Var}(\mathbf{X}_w) > 0$. We say that $R(\theta, \mathbf{h})$ has a unique minimizer in $\mathbf{h}$ if, whenever $(\bar{\theta}, \bar{\mathbf{h}})$ and $(\tilde{\theta}, \tilde{\mathbf{h}})$ are minimizers of $R(\theta, \mathbf{h})$, we have $\bar{\mathbf{h}} = \tilde{\mathbf{h}}$.

PROPOSITION 3.3. *Suppose that* (12) *holds and that* $\boldsymbol{\sigma}_w = 0$ *if and only if* $\boldsymbol{\eta}_w = 0$, $w = 1, \ldots, c$. *Then any one of the following three conditions implies* $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} = \mathrm{Span}(\boldsymbol{\beta}_w)$:

1. $\mu_w(t)$ *is a linear function of* t.
2. *Condition* (15) *holds for some* ϕ_w, *for which* $R(a, \mathbf{h})$ *has a unique minimizer in* $\mathbf{h}$, *and* $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\beta}_w'\mathbf{X}_w)$ *is linear in* $\boldsymbol{\beta}_w'\mathbf{X}_w$.
3. *The conditional expectation* $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\eta}_w'\mathbf{X}_w)$ *is linear in* $\boldsymbol{\eta}_w'\mathbf{X}_w$.

In this proposition, (12) is simply that all subpopulation mean functions depend on at most one linear combination of the predictors. All of the location regressions discussed in Section 2.2 are of this form except (10), which allows the mean functions to depend on multiple linear combinations. Many other regressions are of this form as well. Nevertheless, (12) represents a potential limitation of using OLS which cannot, in general, describe regressions with multidirectional subpopulation mean functions. It must be remarked, though, that by combining the spans of subpopulation OLS vectors, we may very well recover the whole CPMS even when $\mathrm{Span}(\boldsymbol{\beta}_w)$ is a proper subset of $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ for some w. Loosely speaking, this is due to the fact that the "difference" $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} \setminus \mathrm{Span}(\boldsymbol{\beta}_w)$ may be recovered along OLS directions of other subpopulations; $\boldsymbol{\beta}_{\tilde{w}}$, $\tilde{w} \neq w$.

The second assumption, $\boldsymbol{\sigma}_w = 0$ if and only if $\boldsymbol{\eta}_w = 0$, is intended to rule out symmetric mean functions for which $\boldsymbol{\sigma}_w = 0$ while $\boldsymbol{\eta}_w \neq 0$. For example, this combination happens when $\mathbf{X}_w$ is a standard normal random vector and

$Y = X_{w1}^2 + \varepsilon$. Generally, the mean functions $\mathrm{E}(Y_w|\mathbf{X}_w)$ must have a linear trend for OLS to be useful.

Proposition 3.3 provides a flexible spectrum of conditions for (13). Proposition 3.3.1 essentially assumes the conclusion. Nevertheless, the statement is useful because it reminds us that the desired conclusion is trivially true if OLS recovers $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ in the population. When (15) holds but $\mu_w(t)$ is not linear in t, such as would be the case if $\mu_w(t)$ were monotone, Proposition 3.3.2 says we can still achieve the desired conclusion by adding the predictor condition that $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\beta}_w'\mathbf{X}_w)$ is linear. Finally, when (15) does not hold, Proposition 3.3.3 says we can still achieve the desired conclusion with the requirement that $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\eta}_w'\mathbf{X}_w)$ be linear.

The predictor linearity condition of Proposition 3.3.3 is similar to the linearity condition of Proposition 3.3.2, but there is one very important difference. While the latter can be checked easily in practice by plotting the individual predictors against the OLS fitted values, the former cannot be checked directly prior to the estimation of $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ itself. Also note that Propositions 3.1 and 3.2 make no reference to the underlying probability or likelihood, except to the extent that the conditional mean derived therefrom is monotone. Thus the context to which they apply is much wider than generalized linear models.

The same linearity as that assumed in Proposition 3.3.3 was postulated in Li and Duan (1989). The conclusion of Proposition 3.3.3, as applied to a single category w, was proved in Li and Duan (1989) for the CS and was proved in Cook and Li (2002) for the CMS. While the proof of Proposition 3.3.2 to some degree echoes that of Theorem 2.1 of Li and Duan (1989) for a single category w, the two results differ in both their assumptions and their conculsions: roughly speaking, the former asserts that the minimizer in $\mathbf{h}$ of $R(a, \mathbf{h})$ is parallel to $\boldsymbol{\beta}_w$ if $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\beta}_w'\mathbf{X}_w)$ is linear in $\boldsymbol{\beta}_w'\mathbf{X}_w$, whereas the latter asserts that the minimizer in $\mathbf{h}$ of $R(a, \mathbf{h})$ is parallel to $\boldsymbol{\eta}_w$ if $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\eta}_w\mathbf{X}_w)$ is linear in $\boldsymbol{\eta}_w'\mathbf{X}_w$. It is this distinction, combined with Propositions 3.1 and 3.2, that offers the mentioned additional flexibility.

Under the conditions we have established in Proposition 3.3 that insure $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} = \mathrm{Span}(\boldsymbol{\beta}_w)$, $w = 1, \ldots, c$, the sample estimator of $\boldsymbol{\beta}_w$ within each group is a consistent estimator of $\boldsymbol{\beta}_w$. However, such individual estimates do not provide information about the interrelationship among them, that is, whether some of them are essentially linearly dependent and whether there is a ranking in significance among them. We now turn to inference for the dimension d of $\mathrm{Span}(\mathbf{B}) = \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}^{(W)}$ that yields such information. We provide methodology for this by developing, in the next section, a means to test a null hypothesis of the form $\mathrm{rank}(\mathbf{B}) = m$ versus the alternative that $\mathrm{rank}(\mathbf{B}) > m$. An estimate of $d = \mathrm{rank}(\mathbf{B})$ is constructed by performing a series of tests: beginning with $m = 0$, test $\mathrm{rank}(\mathbf{B}) = m$ versus $\mathrm{rank}(\mathbf{B}) > m$. If the hypothesis is rejected, increment m by 1 and test again. If the hypothesis is not rejected, the value of m under the current hypothesis is taken as the estimate of d.

4. Large sample tests for rank(B). The rank d of $\mathbf{B}$ does not change if we multiply its column vectors by nonzero scalars, or if we pre- or postmultiply $\mathbf{B}$ by full rank matrices. However, transformations of this kind simplify the derivation of our large sample test. In particular, we will transform $\mathbf{B}$ in such a way that, under appropriate conditions, the relevant asymptotic distribution is chi-squared.

Let $a_w = \{\Pr(W = w)\}^{1/2} > 0$ and define the $p \times c$ matrix $\mathbf{B}^* = (a_1\boldsymbol{\beta}_1, \ldots, a_c\boldsymbol{\beta}_c)$. Next, we define the average within subpopulation covariance matrix,

$$\boldsymbol{\Sigma}_\cdot = \sum_{w=1}^{c} a_w^2 \boldsymbol{\Sigma}_w = \mathrm{E}(\boldsymbol{\Sigma}_W),$$

and assume that it is positive definite. Last, we consider the subpopulation least square residuals

$$e_w = \big(Y_w - \mathrm{E}(Y_w)\big) - \boldsymbol{\beta}_w'\big(\mathbf{X}_w - \mathrm{E}(\mathbf{X}_w)\big), \qquad w = 1, \ldots, c, \tag{17}$$

with their variances $\omega_w = \mathrm{Var}(e_w) > 0$ arranged in the $c \times c$ diagonal matrix $\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_c)$. We can now replace the hypothesis $\mathrm{rank}(\mathbf{B}) = m$ with the equivalent

$$H_0 : \mathrm{rank}(\boldsymbol{\Sigma}_\cdot^{1/2}\mathbf{B}^*\boldsymbol{\Omega}^{-1/2}) = m \tag{18}$$

and construct sample estimates of the matrices involved in this null hypothesis.

For each $w = 1, \ldots, c$, let $\{(\mathbf{X}_{iw}, Y_{iw}) : i = 1, \ldots, n_w\}$ be independent observations of $(\mathbf{X}, Y)$ from subpopulation w and let $\bar{Y}_w$ and $\bar{\mathbf{X}}_w$ be the corresponding sample averages. The quantities $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\sigma}_w$ will be estimated by

$$\hat{\boldsymbol{\Sigma}}_w = \frac{1}{n_w}\sum_{i=1}^{n_w}(\mathbf{X}_{iw} - \bar{\mathbf{X}}_w)(\mathbf{X}_{iw} - \bar{\mathbf{X}}_w)'$$

and

$$\hat{\boldsymbol{\sigma}}_w = \frac{1}{n_w}\sum_{i=1}^{n_w}(\mathbf{X}_{iw} - \bar{\mathbf{X}}_w)(Y_{iw} - \bar{Y}_w).$$

The subpopulation OLS vectors will be estimated by $\hat{\boldsymbol{\beta}}_w = \hat{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\sigma}}_w$ and the subpopulation square-root probabilities will be estimated by $\hat{a}_w = \sqrt{n_w/n}$. We also set $\hat{\mathbf{B}}^* = (\hat{a}_1\hat{\boldsymbol{\beta}}_1, \ldots, \hat{a}_c\hat{\boldsymbol{\beta}}_c)$ and

$$\hat{\boldsymbol{\Sigma}}_\cdot = \sum_{w=1}^{c}\hat{a}_w^2\hat{\boldsymbol{\Sigma}}_w = \frac{1}{n}\sum_{w=1}^{c} n_w\hat{\boldsymbol{\Sigma}}_w.$$

Last, we form the sample residuals

$$\hat{e}_{iw} = (Y_{iw} - \bar{Y}_w) - \hat{\boldsymbol{\beta}}_w'(\mathbf{X}_{iw} - \bar{\mathbf{X}}_w),$$

estimate the residual variances by

$$\hat{\omega}_w = \frac{1}{n_w}\sum_{i=1}^{n_w}\hat{e}_{iw}^2$$

and construct the matrix $\hat{\mathbf{\Omega}} = \operatorname{diag}(\hat{\omega}_1, \ldots, \hat{\omega}_c)$. By straightforward application of the weak law of large numbers we have that $\hat{\mathbf{B}}^*$, $\hat{\mathbf{\Sigma}}_\cdot$ and $\hat{\mathbf{\Omega}}$ converge in probability to $\mathbf{B}^*$, $\mathbf{\Sigma}_\cdot$ and $\mathbf{\Omega}$, respectively.

Our next objective is to construct a test statistic for (18) based on these sample matrices and to derive its asymptotic distribution. Consider the singular value decomposition

$$\mathbf{\Sigma}_\cdot^{1/2}\mathbf{B}^*\mathbf{\Omega}^{-1/2} = (\tilde{\mathbf{\Gamma}} \quad \mathbf{\Gamma}) \begin{pmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{\Psi}}' \\ \mathbf{\Psi}' \end{pmatrix}, \tag{19}$$

where $(\tilde{\mathbf{\Gamma}} \quad \mathbf{\Gamma})$ is a $p \times p$ orthogonal matrix (left singular vectors), $(\tilde{\mathbf{\Psi}} \quad \mathbf{\Psi})$ is a $c \times c$ orthogonal matrix (right singular vectors) and $\mathbf{D}$ is a $d \times d$ diagonal matrix with nonzero diagonal elements (nonzero singular values). Correspondingly, $\mathbf{\Gamma}$ has dimension $p \times (p-d)$ and $\mathbf{\Psi}$ has dimension $c \times (c-d)$.

The rank of $\mathbf{\Sigma}_\cdot^{1/2}\mathbf{B}^*\mathbf{\Omega}^{-1/2}$ corresponds to the number of nonzero singular values in (19). Thus, a natural test statistic for (18) is

$$T(m) = \sum_{j=m+1}^{p} \hat{\lambda}_j, \tag{20}$$

where $\hat{\lambda}_{m+1} \geq \cdots \geq \hat{\lambda}_p$ are the smallest $p-m$ eigenvalues of the matrix

$$n\big(\hat{\mathbf{\Sigma}}_\cdot^{1/2}\hat{\mathbf{B}}^*\hat{\mathbf{\Omega}}^{-1/2}\big)\big(\hat{\mathbf{\Sigma}}_\cdot^{1/2}\hat{\mathbf{B}}^*\hat{\mathbf{\Omega}}^{-1/2}\big)'$$

or, equivalently, the squares of the $p-m$ singular values of

$$\sqrt{n}\hat{\mathbf{\Sigma}}_\cdot^{1/2}\hat{\mathbf{B}}^*\hat{\mathbf{\Omega}}^{-1/2}$$

that are closest to 0.

To use $T(m)$ in practice, we need its asymptotic distribution under the hypothesis $d = m$. By Eaton and Tyler (1994), the joint asymptotic distribution of these $p-m$ singular values is the same as that of the singular values of the matrix

$$\mathbf{U} = \sqrt{n}\mathbf{\Gamma}'\big(\hat{\mathbf{\Sigma}}_\cdot^{1/2}\hat{\mathbf{B}}^*\hat{\mathbf{\Omega}}^{-1/2} - \mathbf{\Sigma}_\cdot^{1/2}\mathbf{B}^*\mathbf{\Omega}^{-1/2}\big)\mathbf{\Psi}.$$

The next propositions concern the asymptotic distribution of $\mathbf{U}$, or rather of the random vector $\operatorname{vec}(\mathbf{U})$ obtained by stacking its columns, and the associated asymptotic distribution of $T(d)$. Let $\boldsymbol{\psi}_w'$ be the row vectors of $\mathbf{\Psi}$, let $\mathbf{Z}_w = \mathbf{\Sigma}_w^{-1/2}(\mathbf{X}_w - \mathrm{E}(\mathbf{X}_w))$, $w = 1, \ldots, c$, be the standardized subpopulation predictors and let $\otimes$ indicate the Kronecker product.

PROPOSITION 4.1. *Assuming that all moments involved are finite, then the random vector* $\operatorname{vec}(\mathbf{U})$ *has an asymptotic* $(p-d)(c-d)$*-dimensional multivariate*

normal distribution with mean $\mathbf{0}$ *and covariance matrix*

$$\begin{aligned}\boldsymbol{\Delta} = \sum_{w=1}^{c} (\omega_w^{-1/2}\boldsymbol{\psi}_w \otimes \boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Sigma}_w^{-1/2}) \\ \times \mathrm{E}(e_w^2\mathbf{Z}_w\mathbf{Z}_w')(\omega_w^{-1/2}\boldsymbol{\psi}_w' \otimes \boldsymbol{\Sigma}_w^{-1/2}\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Gamma}).\end{aligned} \tag{21}$$

Thus

$$T(d) \xrightarrow{\mathcal{L}} \sum_{j=1}^{(p-d)(c-d)} \alpha_j K_j,$$

where $\alpha_1, \ldots, \alpha_{(p-d)(c-d)}$ *are the eigenvalues of* (21) *and* $K_1, \ldots, K_{(p-d)(c-d)}$ *are independent* χ_1^2 *random variables*.

There are two ways the results of this proposition might be used in practice. First, under the hypothesis that $d = m$, obtain estimates $\hat{\alpha}_j$ of the eigenvalues α_j from a sample version $\hat{\boldsymbol{\Delta}}$ of $\boldsymbol{\Delta}$ constructed by substituting moment estimates for unknown parameters. Then refer $T(m)$ to its estimated asymptotic distribution to obtain a p-value.

Alternatively, instead of calculating percentage points for the distribution of a combination of chi-squares, we can resort to an approach employed with satisfactory results by Bentler and Xie (2000) in the context of PHD. In this approach, estimates of the mean and variance of $T(m)$ under the hypothesis $d = m$ are used to adjust $T(m)$ to yield a statistic that has null distributions closer to a chi-square. The *adjusted statistic* proposed by Satterthwaite (1941),

$$\tilde{T}(m) = \frac{r}{\operatorname{trace}(\hat{\boldsymbol{\Delta}})} T(m),$$

where r is the closest integer to $\operatorname{trace}^2(\hat{\boldsymbol{\Delta}})/\operatorname{trace}(\hat{\boldsymbol{\Delta}}^2)$, can be compared to the percentage points of a χ_r^2. In a simulation study, Fouladi (1997) found the adjusted statistic to perform better than competitors.

We next consider situations in which the asymptotic distribution of $T(d)$ simplifies.

COROLLARY 4.1. *Suppose that*

$$\operatorname{Cov}(e_w^2, \mathbf{Z}_w\mathbf{Z}_w') = \mathbf{0}_{p\times p}, \qquad w = 1, \ldots, c. \tag{22}$$

Then $\boldsymbol{\Delta}$ *defined in* (21) *reduces to*

$$\boldsymbol{\Delta} = \sum_{w=1}^{c} (\boldsymbol{\psi}_w\boldsymbol{\psi}_w') \otimes (\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Gamma}). \tag{23}$$

If, in addition,

$$\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_c, \tag{24}$$

then $\mathbf{\Delta}$ *in* (23) *further simplifies to* $\mathbf{I}_{(p-d)(c-d)}$, *so that*

$$T(d) \xrightarrow{\mathcal{L}} \chi^2_{(p-d)(c-d)}.$$

The first part of the corollary follows because (22) implies

$$\mathrm{E}(e_w^2 \mathbf{Z}_w \mathbf{Z}_w') = \mathrm{E}(e_w^2)\mathrm{E}(\mathbf{Z}_w \mathbf{Z}_w') = \omega_w \mathbf{I} \qquad \text{for } w = 1, \ldots, c.$$

Condition (22) holds whenever the subpopulation linear regression models from which the e_w's are computed are "true" so that $e_w \perp\!\!\!\perp \mathbf{X}_w$. Moreover, as we will see in the next proposition, it holds if $\mathbf{X}_w$ is normal and $\boldsymbol{\beta}_w$ spans the space $\mathcal{S}_{Y_w|\mathbf{X}_w}$. Under the additional assumption of constant conditional covariance matrices (24), we further simplify $\mathbf{\Delta}$ to the identity in the second part of the corollary and thus obtain a chi-square distribution for $T(d)$. The common covariance condition (24) should be reasonable in many regressions.

There is a close connection between condition (22) and PHD. Consider applying PHD to the subpopulation regression of e_w^2 on $\mathbf{X}_w$. In this application of PHD we estimate the rank of the matrix $\mathbf{\Sigma}_{e_w^2 \mathbf{Z}\mathbf{Z}} = \mathrm{E}\{(e_w^2 - \mathrm{E}(e_w^2))\mathbf{Z}_w \mathbf{Z}_w'\}$. Because $\mathbf{\Sigma}_{e_w^2 \mathbf{Z}\mathbf{Z}} = 0$ if and only if (22) holds, we can use PHD straightforwardly to test (22) by testing $\mathrm{rank}(\mathbf{\Sigma}_{e_w^2 \mathbf{Z}\mathbf{Z}}) = 0$ within each subpopulation. This type of application of PHD was discussed by Cook [(1998b), Section 6.2].

The distinction between $d = 0$ and $d > 0$ may be important in diagnostic investigations where Y is a residual. When $d = 0$, $\mathbf{\Delta}$ reduces to a $c \times c$ matrix of $p \times p$ blocks. The off diagonal blocks are all zero and the wth diagonal block is $\mathbf{\Sigma}_{\cdot}^{1/2} \mathbf{\Sigma}_w^{-1} \mathbf{\Sigma}_{\cdot}^{1/2}$.

The next proposition gives sufficient conditions for (22).

PROPOSITION 4.2. *Assume* (12). *Condition* (22) *holds if, for each subpopulation* w,

1. $Y_w \perp\!\!\!\perp \mathbf{X}_w | \boldsymbol{\beta}_w' \mathbf{X}_w$ *and*
2. $\mathbf{X}_w$ *is normally distributed for* $w = 1, \ldots, c$.

As is often the case in the theory of sufficient dimension reduction, this proposition allows us to weaken requirements on the regression structure by strengthening requirements on the predictors: If $\mathbf{X}_w$ is normal, instead of justifying (22) assuming the model underlying the e_w residual to be true, we can justify it assuming the regression of Y_w on $\mathbf{X}_w$ to be a location regression because under (13) this is equivalent to condition 1. Passing from (21) to (22) is useful in practice, because through the latter, the estimation of $\mathbf{\Delta}$ involves only second-order moments.

5. Computational summary. We now summarize the computations involved in estimating $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$ using OLS and constructing the corresponding summary plot.

- Compute $\hat{\boldsymbol{\Sigma}}_{\cdot}$, $\hat{\mathbf{B}}^*$ and $\hat{\boldsymbol{\Omega}}$ from the data, and calculate the spectral decomposition

$$n\big(\hat{\boldsymbol{\Sigma}}_{\cdot}^{1/2}\hat{\mathbf{B}}^*\hat{\boldsymbol{\Omega}}^{-1/2}\big)\big(\hat{\boldsymbol{\Sigma}}_{\cdot}^{1/2}\,\hat{\mathbf{B}}^*\,\hat{\boldsymbol{\Omega}}^{-1/2}\big)' = \sum_{j=1}^{p}\hat{\lambda}_j\hat{\boldsymbol{\gamma}}_j\hat{\boldsymbol{\gamma}}_j' \tag{25}$$

  with eigenvalues $\hat{\lambda}_j$ in nonincreasing order.
- Form the statistics

$$T(m) = \sum_{j=m+1}^{p}\hat{\lambda}_j, \qquad m = 0, \ldots, p-1,$$

  and test sequentially

$$H_0 : \operatorname{rank}\big(\boldsymbol{\Sigma}_{\cdot}^{1/2}\mathbf{B}^*\boldsymbol{\Omega}^{-1/2}\big) = m,$$
$$H_1 : \operatorname{rank}\big(\boldsymbol{\Sigma}_{\cdot}^{1/2}\mathbf{B}^*\boldsymbol{\Omega}^{-1/2}\big) > m$$

  employing an appropriate null distribution. Take $\hat{d}$ to be the smallest m for which the null hypothesis is not rejected.
- Letting $\mathbf{s}_j = \hat{\boldsymbol{\Sigma}}_{\cdot}^{-1/2}\hat{\boldsymbol{\gamma}}_j$, construct a plot of Y versus the *sufficient predictors* $\mathbf{s}_j'\mathbf{X}$, $j = 1, \ldots \hat{d}$, with points marked to indicate the W subpopulation. This is the estimated central partial mean view. The space $\operatorname{Span}(\mathbf{s}_1, \ldots, \mathbf{s}_{\hat{d}})$ provides an estimate of $\operatorname{Span}(\mathbf{B})$, since

$$\operatorname{Span}\big(\boldsymbol{\Sigma}_{\cdot}^{1/2}\mathbf{B}^*\boldsymbol{\Omega}^{-1/2}\big) = \boldsymbol{\Sigma}_{\cdot}^{1/2}\operatorname{Span}\big(a_1\omega_1^{-1/2}\boldsymbol{\beta}_1, \ldots, a_c\omega_c^{-1/2}\boldsymbol{\beta}_c\big)$$

  with $a_w\omega_w^{-1/2} \neq 0$, $w = 1, \ldots, c$.

The results of Propositions 3.3, 4.1, 4.2 and Corollary 4.1 can be combined in various ways to match the regression under study. For example, suppose we conclude that for each subpopulation the regression of Y_w on $\mathbf{X}_w$ is described by the additive error model (7). The predictor $\mathbf{X}$ may contain functionally related terms like quadratics and cross products. From Proposition 3.3.1, $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} = \operatorname{Span}(\boldsymbol{\beta}_w)$ and OLS can be used to estimate the central partial mean subspace. In addition, under the subpopulation model, $e_w \perp\!\!\!\perp \mathbf{X}_w$ and consequently (22) holds. This means that the asymptotic distribution of $T(d)$ is a linear combination of chi-squares with coefficients given by the eigenvalues of the version of $\boldsymbol{\Delta}$ in (23). If the regression resulted from a designed experiment with treatments W randomly assigned to experimental units characterized by $\mathbf{X}$, then the common covariance condition (24) holds and the reference distribution is the chi-squared stated in Corollary 4.1.

6. Pooled estimators for increased accuracy. The method developed in the previous sections is based on separate linear regressions within each subpopulation w. If we know that the dimension d of the CPMS is smaller than c and p, however, we can achieve greater accuracy by pooling all observations for the linear regressions. In this section we introduce such a procedure.

If $d < \min(c, p)$, then, under any one of the sufficient conditions of Proposition 3.3, at most d among the vectors $\{\boldsymbol{\beta}_w\}$ are linearly independent, because all of them must belong to $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$. Hence, even though each $\boldsymbol{\beta}_w$ is obtained by minimizing $\mathrm{E}(Y_w - \theta_w - \mathbf{h}_w'\mathbf{X}_w)^2$ separately, the fact that $d < \min(c, p)$ forces linear dependencies among these vectors. So there are a $p \times d$ matrix $\bar{\mathbf{H}}$ and d-dimensional vectors $\bar{\boldsymbol{\rho}}_1, \ldots, \bar{\boldsymbol{\rho}}_c$ such that $\boldsymbol{\beta}_w = \bar{\mathbf{H}}\bar{\boldsymbol{\rho}}_w$ for all $w = 1, \ldots, c$. This implies that the minimization of

$$\mathrm{E}\big[(Y - \theta_W - \mathbf{h}_W'\mathbf{X})^2/\omega_W\big]$$

over $\{\theta_w\}$ and $\{\mathbf{h}_w\}$ is equivalent to the minimization of

$$\mathrm{E}\big[(Y - \theta_W - \boldsymbol{\rho}_W'\mathbf{H}'\mathbf{X})^2/\omega_W\big]$$

over all scalars $\{\theta_w\}$, $p \times d$ matrices $\mathbf{H}$ and $d \times 1$ vectors $\{\boldsymbol{\rho}_w\}$, even though the latter is carried out on a smaller parameter space. The next proposition summarizes this fact.

PROPOSITION 6.1. *Suppose* (a) $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$ *has dimension* $d < \min(c, p)$ *with basis* $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d)$, (b) *each* $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ *has dimension* 1 *and* (c) *for each* w, $\mathrm{E}(\mathbf{X}_w|\boldsymbol{\eta}_w'\mathbf{X}_w)$ *is linear in* $\boldsymbol{\eta}_w'\mathbf{X}_w$. *Then* $(\bar{\mathbf{H}}\bar{\boldsymbol{\rho}}_1, \ldots, \bar{\mathbf{H}}\bar{\boldsymbol{\rho}}_c)$ *spans the space* $\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}$ *as long as all of these vectors are nonzero.*

We now replace d with $\hat{d}$, the estimate of d from Section 5, to construct the pooled estimator of the CPMS. Instead of doing an ordinary least squares fit separately for each subpopulation, we minimize the joint objective function

$$\sum_{w=1}^{c} \hat{\omega}_w^{-1} \sum_{i=1}^{n_w} (Y_{iw} - \theta_w - \boldsymbol{\rho}_w'\mathbf{H}'\mathbf{X}_{iw})^2$$

over all $\{\theta_w\}$, $\mathbf{H}$ and $\{\boldsymbol{\rho}_w\}$. Letting $\{\hat{\theta}_w\}$, $\widehat{\mathbf{H}}$ and $\{\hat{\boldsymbol{\rho}}_w\}$ be the minimizers, we use $\hat{\boldsymbol{\beta}}_w = \widehat{\mathbf{H}}\hat{\boldsymbol{\rho}}_w$ to estimate $\boldsymbol{\beta}_w$. This minimization is not equivalent to the separate OLS fits, because it pools all the data to estimate $\mathbf{H}$ and $\boldsymbol{\rho}_w$, which has a smaller dimension than $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_c)$ if $\hat{d} < \min(c, p)$. If $\hat{d} \geq \min(c, p)$, then this method reduces to the separate OLS fits.

As an illustration we consider the estimation of a CPMS with dimension $d = 1$. Location regressions (6)–(9) and (11) are all of this form when $\mathrm{Span}(\boldsymbol{\eta})$ and

$\operatorname{Span}(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_c)$ each have dimension 1. In this case $\boldsymbol{\rho}_w$ is a scalar, say ρ_w, and $\mathbf{H}$ is a vector, say $\mathbf{h}$, and we are to minimize

$$\sum_{w=1}^{c} \hat{\omega}_w^{-1} \sum_{i=1}^{n_w} (Y_{iw} - \theta_w - \rho_w \mathbf{h}' \mathbf{X}_{iw})^2.$$

By construction, the parameters $\mathbf{h}$ and $\{\rho_w\}$ are not uniquely defined: We can always multiply ρ_w by a constant C and divide $\mathbf{h}$ by C without changing $\rho_w \mathbf{h}$. For uniqueness we impose an arbitrary constraint. For example, we could let $\rho_1 = 1$ and use Newton–Raphson on the remaining parameters or let $\|\mathbf{h}\| = 1$ and use the Lagrangian multiplier method for constrained minimization.

Further development of the pooled estimator regarding its efficiency and computation is beyond the scope of this article and will be tackled in separate research. We provide here a small simulation study to compare, under different circumstances, the pooled estimator and the unpooled estimator discussed in the previous sections.

For simplicity, we consider the case where $p = c = 2$ and $d = 1$. For a given total sample size n, the sample sizes n_1 for the first category are generated from a binomial $(n, 0.5)$ and n_2 is then taken to be $n - n_1$. Once n_1 and n_2 are chosen, we generate the errors ε_{iw}, for $i = 1, \ldots, n_w$ and $w = 1, 2$ from $N(0, 1)$, and generate X_{iw} from the two bivariate normal distributions

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) \quad \text{and} \quad N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \alpha\\\alpha & 1\end{pmatrix}\right).$$

The responses Y_{iw} are generated from the nonlinear model

$$Y_{iw} = \exp\bigl((X_{1i} + X_{2i})/2\bigr) + \varepsilon_{iw}.$$

Thus the CPMS has dimension 1 and is spanned by the vector $(1, 1)'$.

For the comparison we need a measure of error for the estimation of a linear subspace, rather than a specific vector. Evidently any such measure should be length-invariant; that is, it should be affected by the direction but not the length of the estimator. One reasonable choice is as follows. Let $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_{(1)}, \tilde{\beta}_{(2)})'$ be the unpooled estimator and let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_{(1)}, \hat{\beta}_{(2)})'$ be the pooled estimator (here we have used parentheses in the subscript so as not to be confused with the bare index w that denotes category). We define

$$\begin{aligned}
\mathrm{MSE}_1 &= \mathrm{E}\bigl[\bigl(1/\sqrt{2} - \tilde{\beta}_{(1)}/\|\tilde{\boldsymbol{\beta}}\|\bigr)^2 + \bigl(1/\sqrt{2} - \tilde{\beta}_{(2)}/\|\tilde{\boldsymbol{\beta}}\|\bigr)^2\bigr],\\
\mathrm{MSE}_2 &= \mathrm{E}\bigl[\bigl(1/\sqrt{2} - \hat{\beta}_{(1)}/\|\hat{\boldsymbol{\beta}}\|\bigr)^2 + \bigl(1/\sqrt{2} - \hat{\beta}_{(2)}/\|\hat{\boldsymbol{\beta}}\|\bigr)^2\bigr].
\end{aligned}$$

Note that every vector involved is first rescaled to have length 1.

We will make the comparison for varied total sample sizes n and varied correlation α, for we expect the comparison to be affected by the sample size and the difference among the covariance matrices $\boldsymbol{\Sigma}_w$. Table 1 summarizes the result.

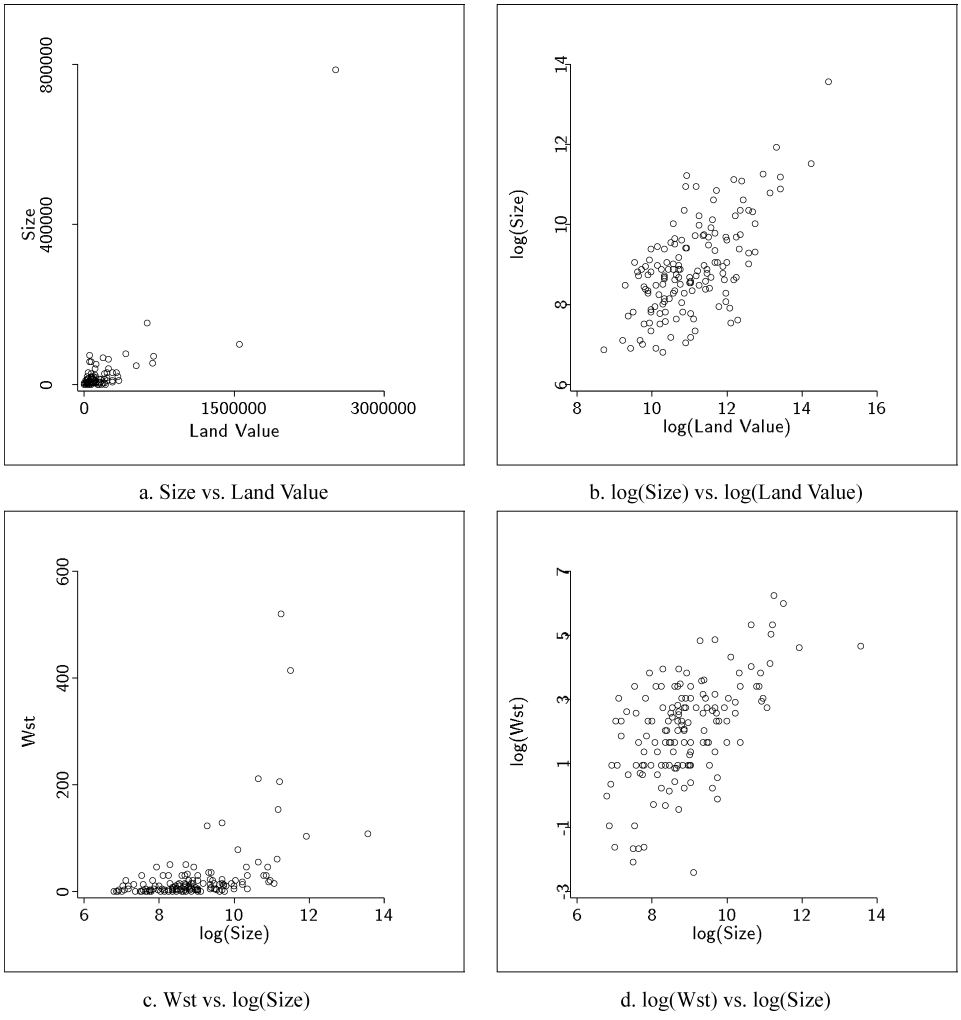
TABLE 1
Comparison of the unpooled and pooled estimators

	$n = 35$		$n = 40$		$n = 45$	
α	$\mathbf{MSE_1}$	$\mathbf{MSE_2}$	$\mathbf{MSE_1}$	$\mathbf{MSE_2}$	$\mathbf{MSE_1}$	$\mathbf{MSE_2}$
0.0	0.0806	0.1016	0.0612	0.0595	0.0449	0.0460
0.1	0.0789	0.0875	0.0626	0.0743	0.0495	0.0459
0.2	0.0757	0.0793	0.0604	0.0593	0.0533	0.0565
0.3	0.0765	0.0821	0.0748	0.0764	0.0552	0.0538
0.4	0.0869	0.0948	0.0777	0.0742	0.0563	0.0495
0.5	0.0917	0.1025	0.0814	0.0776	0.0712	0.0678
	$n = 70$		$n = 80$		$n = 90$	
α	$\mathbf{MSE_1}$	$\mathbf{MSE_2}$	$\mathbf{MSE_1}$	$\mathbf{MSE_2}$	$\mathbf{MSE_1}$	$\mathbf{MSE_2}$
0.0	0.0270	0.0259	0.0236	0.0223	0.0191	0.0184
0.1	0.0282	0.0278	0.0238	0.0231	0.0220	0.0217
0.2	0.0286	0.0276	0.0245	0.0238	0.0219	0.0212
0.3	0.0326	0.0301	0.0302	0.0291	0.0238	0.0222
0.4	0.0370	0.0333	0.0298	0.0272	0.0248	0.0226
0.5	0.0402	0.0344	0.0331	0.0295	0.0278	0.0243

A clear pattern emerges from the table: For small sample sizes ($n = 35$, with each category having about 17 observations), the unpooled estimator works better; for median sample sizes ($n = 40, 45$), the two estimators are about the same (with the unpooled estimator working slightly better) when $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are close, but the pooled estimator works better when $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are different; for larger sample sizes ($n = 70, 80, 90$), the pooled estimator works better and the contrast increases as the difference between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ increases.

7. Applications.

7.1. *Waste tax.* A large urban county in Minnesota needed to add a business tax to support its new waste processing facility. In an effort to be fair and to reduce the potential for lawsuits, the county wanted to tax businesses according to the amount of waste they produce. Businesses do not necessarily keep accurate records on their waste production and, even if they did, the county could not legally compel them to disclose it. The county decided to study the feasibility of using information in the public tax records to predict yearly waste production (Wst) in tons per year. A random sample of 150 businesses was selected from the county records and each was asked to cooperate with county workers to determine their yearly waste production. Three businesses refused, leaving a sample size of 147. The available quantitative predictors from the county's tax records were land value, improved value, structure size in square feet, and number of full time equivalent employees. In addition, information in the public record was used to classify businesses into



a. Size vs. Land Value

b. log(Size) vs. log(Land Value)

c. Wst vs. log(Size)

d. log(Wst) vs. log(Size)

FIG. 1. *Four plots to illustrate selected characteristics of the waste data.*

one of five types (W): manufacturing, warehouse or storage, office building, retail and restaurant or entertainment.

The marginal distribution of each of the four quantitative predictors is highly skewed to the right as illustrated by the plot of size versus land value in Figure 1(a). This indicates that nearly any type of regression analysis would benefit from predictor transformations. We used the likelihood methods implemented in *Arc* [Cook and Weisberg (1999)] to investigate simultaneous power transformations of the four quantitative predictors so that the conditional distribution of the transformed predictors $\mathbf{X}_w$ is approximately normal with common covariance matrix, $w = 1, \ldots, 5$, leading to the log transformation for each of the quantitative

predictors. The result is illustrated by the plot of the logarithms of size and land value in Figure 1(b). This procedure is often effective for insuring that covariances $\mathbf{\Sigma}_w$ are constant and that the subpopulation linearity conditions of Proposition 3.3 are met to an adequate approximation.

Figure 1(c) shows a plot of Wst versus the logarithm of size. A transformation of Wst is clearly indicated. Studying the individual regressions of Wst_w on $\mathbf{X}_w$ we concluded that model (7) would likely hold in terms of the transformed response $Y = \log \text{Wst}$, as illustrated in Figure 1(d).

Assuming that model (7) is accurate, then condition (22) holds and, by Proposition 3.3.1, $\mathcal{S}_{\text{E}(Y_w|\mathbf{X}_w)} = \text{Span}(\boldsymbol{\beta}_w)$. In this setting the distribution of $\mathbf{X}_w$ is used only to determine the reference distribution for the test statistic $T(m)$ as stated in Corollary 4.1. Because the predictor transformations were selected in part to yield constant covariance matrices $\mathbf{\Sigma}_w$, we next applied the dimension reduction method of Section 4 using the chi-square reference distribution. The resulting p-values were 0, 0.45, 0.83 and 0.86 for the hypotheses $d = 0, 1, 2$ and 3. Consequently, we inferred that $\dim(\mathcal{S}^{(W)}_{\text{E}(Y|X)}) = 1$, which is the same as the result obtained by using the adjusted statistic $\tilde{T}$.

The central partial mean view shown in Figure 2 is a plot of the sufficient predictor $\mathbf{s}_1'\mathbf{X}$ (Section 5) with points marked according to business type. The

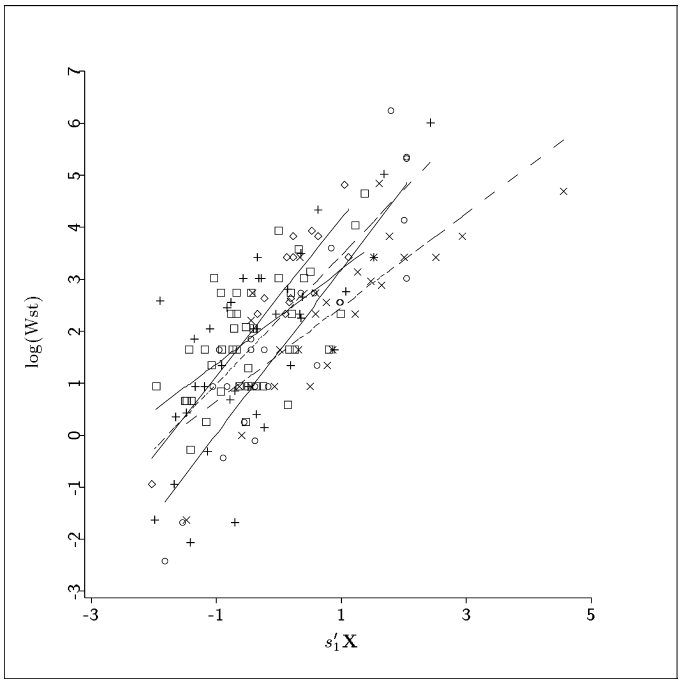


FIG. 2. *Summary plot of the response versus the sufficient predictor* $\mathbf{s}_1'\mathbf{X}$. *Business types are indicated with different plotting symbols. The lines correspond to the linear OLS fits by group.*

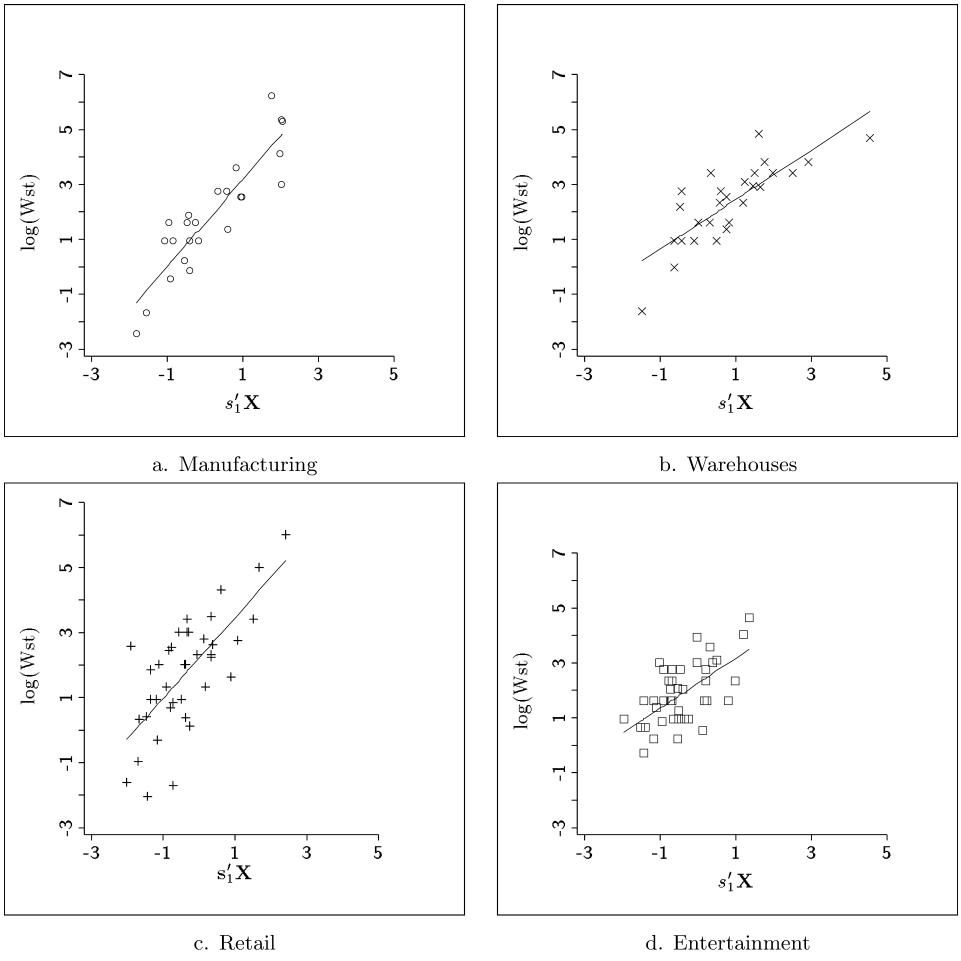


FIG. 3. *Scatter plot of Figure* 2 *with four business types plotted separately.*

subpopulation OLS fits are shown as well. Without interactive features, the plot may be difficult to interpret, so we present in Figure 3 separate plots for four business types. The plots in Figures 2 and 3 together with the inference that $\dim(\mathcal{S}^{(W)}_{\mathrm{E}(Y|X)}) = 1$ suggests that a model of the form

$$Y_w = \mu_w + \alpha_w \boldsymbol{\eta}'\mathbf{X} + \sigma\varepsilon \tag{26}$$

with $\|\boldsymbol{\eta}\| = 1$ may yield predictions that are near the best possible with the available data. This possibility is supported by various standard diagnostics applied to the fitted version of (26) obtained by using OLS. The correlation between the fitted values from (26) and the sufficient predictor $\mathbf{s}_1'\mathbf{X}$ is 0.996, which indicates that the two analyses are finding the same structure in the data. The summary

plots based on the fitted values are visually indistinguishable from those in Figures 2 and 3.

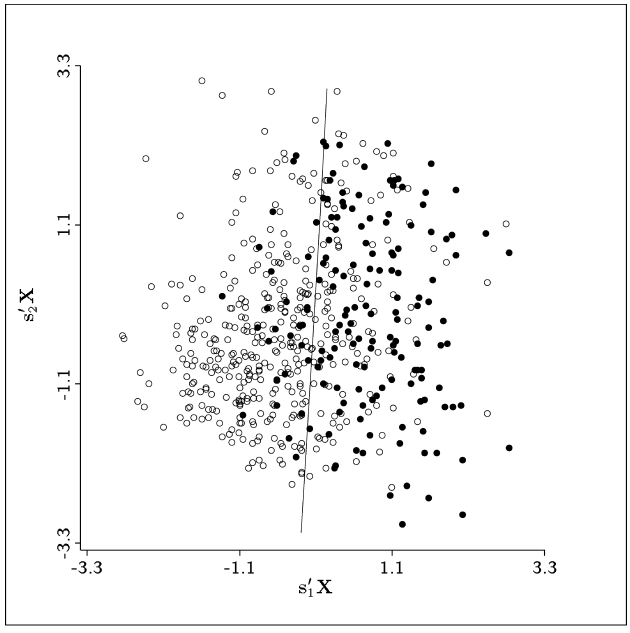
There are other ways the results of this article could be used to study the waste data. Suppose that, after transforming to Y and $\mathbf{X}$, we assume $\dim(\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}) = 1$ without applying the dimension reduction method of Section 4 and without assuming model (26). In this case, Proposition 6.1 is applicable, assuming the predictor linearity condition which seems likely in view of our transformations to multivariate normality. Using the OLS objective function leads back to the summary plots of Figures 2 and 3, again suggesting (26) as a first nonlinear model subject to the usual diagnostic procedures.

Transforming Wst at the outset is clearly sensible. Nevertheless, for illustration, consider the regression of Wst on $(\mathbf{X}, W)$. Because model (7) is no longer appropriate as indicated by Figure 1(c), the distribution of $\mathbf{X}_w$ now takes on added importance. Assuming that $\mathbf{X}_w$ is normally distributed or approximately so, Propositions 3.3 and 4.2 are applicable and we can still use the methods of Section 4. The resulting p-values were 0, 0.29, 0.93 and 0.99 for the hypotheses $d = 0, 1, 2$ and 3, so we inferred that $\dim(\mathcal{S}^{(W)}_{\mathrm{E}(\mathrm{Wst}|\mathbf{X})}) = 1$. The correlation between the sufficient predictors $\boldsymbol{\eta}_1'\mathbf{X}$ from the analyses based on Y and Wst was 0.95. The agreement between these two analyses seems remarkable in view of the extreme skew illustrated in Figure 1(c). The central partial mean view from the analysis based on Wst (not shown) again suggests that the response should be transformed, leading us back to the analysis based on Y.

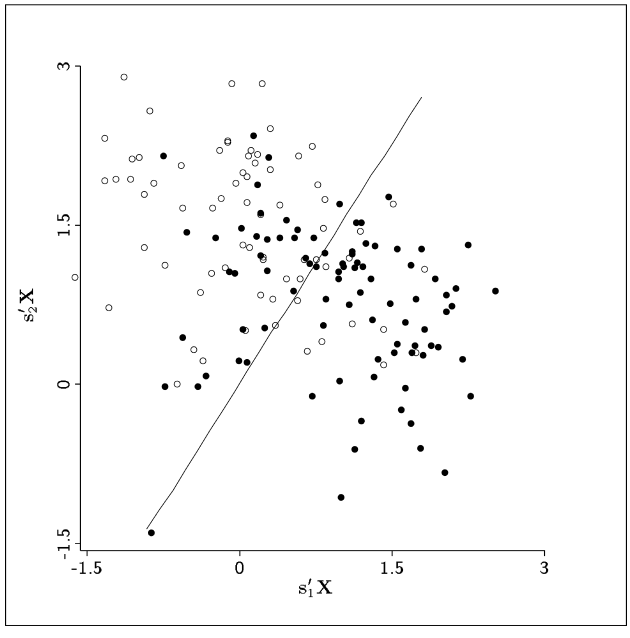
7.2. *Diabetes.* We next consider a data set on 724 female Pima Indian patients, who had complete records from the National Institute of Diabetes and Digestive and Kidney Disease. Smith, Everhart, Dickson, Knowler and Johannes (1988) used this data set to forecast the onset of diabetes mellitus. Cook and Lee (1999) used it to illustrate dimension reduction in regressions with a binary response.

The binary response variable Y equals 1 if a patient tested positive for diabetes and 0 otherwise. The quantitative predictors we considered in $\mathbf{X}$ are diastolic blood pressure, logarithms of the body mass index and the diabetes pedigree function, inverse of patient's age, and cube root of plasma glucose concentration. As a categorical predictor W we considered the number of times the patient was pregnant in $c = 5$ classes: 0, 1 or 2, 3 or 4, 5 or 6, and more than 6. For later reference, we labeled the final class the 7+ class. The class counts for W are 99, 227, 137, 101 and 160. As in the waste regression, the power transformations used to form $\mathbf{X}$ were chosen so that $\mathbf{X}_w$ is approximately multivariate normal with common covariance matrix for $w = 1, \ldots, 5$.

As in the previous illustration, we assume $\dim(\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}) \le 1$, $w = 1, \ldots, 5$, which is supported by our analyses of individual subpopulation regressions. The required linearity condition on the predictors should hold to an adequate approxi-



a. First four classes



b. Class 7+

FIG. 4. *Plots of the first and second sufficient predictors* $\mathbf{s}_2'\mathbf{X}$ *versus* $\mathbf{s}_1'\mathbf{X}$ *from diabetes regression*: (a) *the first four W classes*; (b) *class* 7+. *Lines correspond to Fisher's linear discriminant function in* $(\mathbf{s}_1'\mathbf{X}, \mathbf{s}_2'\mathbf{X})$ *for the data in the respective plots*, ∘ *denotes negative diabetes test and* • *denotes positive diabetes test*.

mation in view of prior normalizing transformations. Similarly, Proposition 3.3.3 implies that $\mathrm{Span}(\boldsymbol{\beta}_w) = \mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ so we can again apply the dimension reduction method of Section 4. Because the response is binary, condition (22), which gives one way to simplify the asymptotic distribution of $T(d)$, does not follow directly from the model. Nevertheless, Proposition 4.2 should be appropriate and, assuming constant covariance matrices $\boldsymbol{\Sigma}_w$, we may use the chi-square reference distribution.

Application of the dimension reduction method in Section 4 to the regression of Y on $(\mathbf{X}, W)$ gave the p-values 0, 0.003 and 0.774 for dimension hypotheses $d = 0, 1$ and 2. We therefore inferred that $\dim(\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}) = 2$, and constructed the sufficient predictors $\mathbf{s}_1'\mathbf{X}$ and $\mathbf{s}_2'\mathbf{X}$. As a first visualization of the results, we used *Arc* [Cook and Weisberg (1999)] to construct a three-dimensional plot (not shown) of Y versus $\mathbf{s}_1'\mathbf{X}$ and $\mathbf{s}_2'\mathbf{X}$. We then superimposed surfaces formed by the fitted probabilities from logistic regressions of Y on $(\mathbf{s}_1'\mathbf{X}, \mathbf{s}_2'\mathbf{X})$ within each of the five classes. A striking feature of the plot was that the fitted probability surface for the 7+ class was very different from the surfaces for the other four classes, which seemed relatively similar.

Because the 7+ class seemed very different from the others, we removed it and recomputed the tests for dimension with the remaining four classes, obtaining p-values of 0 and 0.846 for the hypotheses $d = 0$ and $d = 1$. Thus we inferred that, without the 7+ class, $\dim(\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}) = 1$. The sample correlation between the $\mathbf{s}_1'\mathbf{X}$ predictors with and without the 7+ class was 0.99, indicating that the first estimated direction remains the same with and without 7+.

As a visual check on these inferences, we computed Fisher's linear discriminant function based on $(\mathbf{s}_1'\mathbf{X}, \mathbf{s}_2'\mathbf{X})$, separately for each of the five classes. In the first four classes, $\mathbf{s}_2'\mathbf{X}$ did not add significant information beyond that from $\mathbf{s}_1'\mathbf{X}$. However, in class 7+ both predictors were needed for discrimination. Shown in Figure 4 are plots of $\mathbf{s}_2'\mathbf{X}$ versus $\mathbf{s}_1'\mathbf{X}$ for the combined data from the first four classes and for class 7+. The lines superimposed to the two plots are Fisher's linear discriminant for the combined data from the first four classes and for class 7+. The slope of the line in Figure 4(a) is about 46.8, while that of the line in Figure 4(b) is about 1.2. These results support the inference that $\mathbf{s}_1'\mathbf{X}$ is sufficient for the first four classes, while a different linear combination is needed for class 7+.

There are several ways to continue such an analysis, depending on the application context. For instance, we could perform a separate analysis for class 7+. If one overall model is desired, then it most likely should incorporate interactions between the quantitative predictors and an indicator for class 7+. However the analysis is continued, the finding that for the first four classes $\dim(\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})}) = 1$, while class 7+ behaves differently and induces a second relevant direction, will likely play a fundamental role.

8. Discussion. In this article, we developed methodology for dimension reduction in regressions with categorical predictors, when the objective is to preserve information on the conditional mean. The methodology can be understood and employed on at least three different levels.

First, it can be employed as an alternative or supplement to linear modeling with indicator variables. In the waste data application in Section 7.1, the standard approach leads to a linear regression mean function with 21 terms, including 16 interactions. With our approach, we investigated the applicability of model (7) one category at a time, and then applied dimension reduction. Inferring the dimension of the CPMS to be one leads to summary plots analogous to those in Figures 2 and 3. These can in turn be taken as a starting point for parametric modeling of the mean function across categories. By reducing the burden of rationalizing and selecting among a very large number of terms, this alternative route may greatly facilitate understanding for students in a typical regression service course.

When the dimension of the CPMS is inferred to be larger than one, as in the diabetes data application of Section 7.1, an interesting question concerns individual category contributions to the CPMS (e.g., the second relevant direction required by class 7+). A $c \times c$ scatter plot matrix of $\{\hat{\boldsymbol{\beta}}_w' \mathbf{X}_i\}$, $i = 1, \ldots, n$, with points marked by category, is an effective graphical tool for investigating these contributions. If two $\hat{\boldsymbol{\beta}}_w$'s are nearly parallel, as would be expected when the two categories furnish the same direction to the CPMS, then the points in the corresponding frame of the scatter plot matrix will be highly correlated. Note that the large sample results developed in Section 4 can be straightforwardly adapted for inference on the dimension of the space spanned by a subset of the $\boldsymbol{\beta}_w$'s. Thus, we are in a position to assess individual contributions and compare dimensions spanned with and without given categories.

Second, our methodology can be used in the context of generalized linear models, and thus might be appropriate also in regression courses for statistics majors. For instance, if it is found that a logistic model $\text{logit}(\mathbf{X}_w) = \mu_w + \boldsymbol{\eta}_w' \mathbf{X}_w$ is appropriate within categories, we can still use OLS and $\hat{\mathbf{B}}$ to reduce the dimension provided that $\text{E}(\mathbf{X}_w | \boldsymbol{\beta}_w' \mathbf{X}_w)$ is linear or approximately linear for each w (Proposition 3.3.2). In this setting, it may be possible to improve efficiency by replacing $\hat{\mathbf{B}}$ with the corresponding matrix of coefficient estimates $\hat{\boldsymbol{\eta}}_w$ from intra-category logistic fits, but in view of the exploratory role of the methodology, OLS will usually suffice.

Third, our methodology can be employed with no reference to models within categories. In fact, OLS and $\hat{\mathbf{B}}$ can again be used for dimension reduction, provided that $\text{E}(\mathbf{X}_w | \boldsymbol{\eta}_w' \mathbf{X}_w)$ is linear or approximately linear for each w (Proposition 3.3.3). This type of usage was discussed in both the applications in Section 7.

At all three levels, the approach described in this article can be quite useful in dealing with regression data that involve several quantitative variables and, along with them, categorical ones.

Our methods for the central partial mean subspace rely on the nontrivial assumption that each intracategory CMS has at most dimension one—condition (12). Methodology exists for multidimensional central mean subspaces when no categorical predictors are involved [Cook and Li (2002)], as well as for multidimensional central partial subspaces which encompass categorical predictors but do not restrict attention to the mean [Chiaromonte, Cook and Li (2002)]. We are currently investigating methods for central partial mean subspaces that do not require (12). It must be noted, though, that in comparison to the SIR-based methods in Chiaromonte, Cook and Li (2002), the approach presented here has the advantage of relying less heavily on the assumption of common covariance across categories. In fact, while condition (24) simplifies the asymptotic distribution of $T(d)$, our large sample results allow us to perform dimensional inference through linear combinations of chi-squares or adjusted statistics also when (24) does not hold.

APPENDIX

PROOF OF PROPOSITION 2.1. That (5) implies Proposition 2.1.1 is immediate. That Proposition 2.1.2 implies (5) is also immediate because, if 2.1.2 is true, then $\mathrm{E}(Y|\mathbf{X}, W)$ is constant given $(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W)$ and is therefore independent of Y given $(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W)$. Now let us prove that Proposition 2.1.1 implies Proposition 2.1.2. By Proposition 2.1.1,

$$\mathrm{E}[Y\mathrm{E}(Y|\mathbf{X}, W)|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W] = \mathrm{E}(Y|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W)\mathrm{E}[\mathrm{E}(Y|\mathbf{X}, W)|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W].$$

The left-hand side of this equation is

$$\mathrm{E}[Y\mathrm{E}(Y|\mathbf{X}, W)|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W] = \mathrm{E}[\mathrm{E}^2(Y|\mathbf{X}, W)|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W]$$

and the right-hand side is

$$\mathrm{E}^2[\mathrm{E}(Y|\mathbf{X}, W)|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W].$$

Consequently, $\mathrm{Var}[\mathrm{E}(Y|\mathbf{X}, W)|\mathbf{P}_{\mathcal{S}}\mathbf{X}, W] = 0$, which means that $\mathrm{E}(Y|\mathbf{X}, W)$ is a constant given $(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W)$. □

PROOF OF PROPOSITION 2.2. For convenience, let

$$\mathcal{S}^{(W)}_{\mathrm{E}(Y|\mathbf{X})} = \mathcal{S}_1 \quad \text{and} \quad \bigoplus_{w=1}^{c} \mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)} = \mathcal{S}_2.$$

Note that for any subspace $\mathcal{S}$ of $\mathbb{R}^p$, the following two conditional independences are equivalent:

$$Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X}, W)|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W), \tag{27}$$

$$Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X}, W = w)|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W = w) \qquad \text{for } w = 1, \ldots, c. \tag{28}$$

By the definition of CPMS, (27) will be satisfied if $\mathcal{S}$ is taken to be $\mathcal{S}_1$. Therefore, for each $w = 1, \ldots, c$,

$$Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|(\mathbf{P}_{\mathcal{S}_1}\mathbf{X}, W = w).$$

However, since $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ is the CMS for each subpopulation w, the above conditional independence implies that $\mathcal{S}_1$ contains $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ for all $w = 1, \ldots, c$ and consequently also contains their direct sum $\mathcal{S}_2$.

On the other hand, because $\mathcal{S}_2$ contains each $\mathcal{S}_{\mathrm{E}(Y_w|\mathbf{X}_w)}$ and because the latter is the CMS for the subpopulation w, the conditional independence (28) will be satisfied if $\mathcal{S}$ is taken to be $\mathcal{S}_2$. However, that means the space $\mathcal{S}_2$ is a partial dimension reduction for conditional mean, which must therefore contain the CPMS $\mathcal{S}_1$. $\square$

PROOF OF PROPOSITION 3.1. Since we will always work within a subpopulation w, for convenience we omit the subscript w. Thus all expectations of the form $\mathrm{E}\{f(\mathbf{X}, Y)\}$ stand for the conditional expectation $\mathrm{E}\{f(\mathbf{X}_w, Y_w)\}$, and $\mathbf{X}$, Y, $\boldsymbol{\eta}$ and $\mu(\cdot)$ stand for $\mathbf{X}_w$, Y_w, $\boldsymbol{\eta}_w$ and $\mu_w(\cdot)$. We assume that $\mu(\cdot)$ is a nondecreasing function, which does not lose generality because otherwise we can redefine $\mathbf{X}$ to be the negative of the original $\mathbf{X}$.

Now consider the objective function

$$R(\theta, \mathbf{h}) = -\mathrm{E}\left[\int_{\boldsymbol{\eta}'\mathbf{X}}^{\theta+\mathbf{h}'\mathbf{X}} \{Y - \mu(s)\}\, ds\right], \tag{29}$$

where, again, the subscript w on R is omitted. The proof is done in two steps. First, we show that $R(\theta, \mathbf{h})$ is a form of $\mathrm{E}L(\theta + \mathbf{h}'\mathbf{X}, Y)$ as defined in Section 3. Second, we show that $(0, \boldsymbol{\eta})$ is the unique minimizer of $R(\theta, \mathbf{h})$.

Let c be a fixed constant that is independent of θ and $\mathbf{h}$, and rewrite the integral on the right-hand side of (29) as

$$Y(\theta + \mathbf{h}'\mathbf{X}) - Y(\boldsymbol{\eta}'\mathbf{X}) - \int_{\boldsymbol{\eta}'\mathbf{X}}^{c} \mu(s)\, ds - \int_{c}^{\theta+\mathbf{h}'\mathbf{X}} \mu(s)\, ds.$$

Hence, if we denote $\mathrm{E}[Y(\boldsymbol{\eta}'\mathbf{X}) + \int_{\boldsymbol{\eta}'\mathbf{X}}^{c} \mu(s)\, ds]$ by c_1, then the risk function $R(\theta, \mathbf{h})$ becomes

$$\mathrm{E}\left[-Y(\theta + \mathbf{h}'\mathbf{X}) + c_1 + \int_{c}^{\theta+\mathbf{h}'\mathbf{X}} \mu(s)\, ds\right].$$

Because $\mu(\cdot)$ is nondecreasing, the function $\int_c^t \mu(s)\, ds$, and hence $c_1 + \int_c^t \mu(s)\, ds$, is convex. Now write the latter function as $\phi(t)$ to complete the first step.

Next, by taking iterative expectations, we can write the expectation on the right-hand side of (29) as

$$\mathrm{E}\left[\int_{\boldsymbol{\eta}'\mathbf{X}}^{\theta+\mathbf{h}'\mathbf{X}} \{\mathrm{E}(Y|\mathbf{X}) - \mu(s)\}\, ds\right] = \mathrm{E}\left[\int_{\boldsymbol{\eta}'\mathbf{X}}^{\theta+\mathbf{h}'\mathbf{X}} \{\mu(\boldsymbol{\eta}'\mathbf{X}) - \mu(s)\}\, ds\right]. \tag{30}$$

If $(\theta, \mathbf{h}) = (0, \boldsymbol{\eta})$, then this expectation is 0. By assumption (16), this expectation is smaller than 0 whenever $(\theta, \mathbf{h}) \neq (0, \boldsymbol{\eta})$. Hence $R(\theta, \mathbf{h})$ has the unique minimizer $(0, \boldsymbol{\eta})$. □

PROOF OF PROPOSITION 3.2. (i) Consider, for any t and u, the integral

$$I(t, u) = \int_u^t \{\mu(u) - \mu(s)\}\, ds.$$

It is easy to see that whenever $t \neq u$, $I(u, t) < 0$. This is because if $u < t$, then $\mu(u) - \mu(s) < 0$ for all s between u and t, and if $u > t$, then $\mu(u) - \mu(s) > 0$ for all s between t and u. In this notation the integral inside the probability in (16) can be written as $I(\boldsymbol{\eta}'\mathbf{X},\ a + \mathbf{h}'\mathbf{X})$. So (16) will hold if

$$\Pr(\theta + \mathbf{h}'\mathbf{X} \neq \boldsymbol{\eta}'\mathbf{X}) > 0. \tag{31}$$

If $\mathbf{h} = \boldsymbol{\eta}$, then $a \neq 0$ and so $a + \mathbf{h}'\mathbf{x} \neq \boldsymbol{\eta}'\mathbf{x}$ holds for all $\mathbf{x}$. If $\mathbf{h} \neq \boldsymbol{\eta}$, then, because $\operatorname{Var}(\mathbf{X})$ is positive definite, we always have

$$\operatorname{Var}(\theta + \mathbf{h}'\mathbf{X} - \boldsymbol{\eta}'\mathbf{X}) > 0.$$

Hence the difference $\theta + \mathbf{h}'\mathbf{X} - \boldsymbol{\eta}'\mathbf{X}$ cannot be a degenerate random variable, which implies (31).

(ii) Since the support $\mathcal{X}$ of $\mathbf{X}$ is an open set in $\mathbb{R}^p$, its intersection with any $(p-1)$-dimensional hyperplane has probability zero. Hence, whenever $(\theta, \mathbf{h}) \neq (0, \boldsymbol{\eta})$,

$$\Pr\left(\theta + \mathbf{h}'\mathbf{X} \neq \boldsymbol{\eta}'\mathbf{X}\right) = 1.$$

In view of this, we can assume without loss of generality that $\theta + \mathbf{h}'\mathbf{x} \neq \boldsymbol{\eta}'\mathbf{x}$ for all $\mathbf{x}$ in $\mathcal{X}$. Furthermore, because $\mathcal{X}$ is convex, the set $\{\boldsymbol{\eta}'\mathbf{x} : \mathbf{x} \in \mathcal{X}\}$ is an interval, say I, and, for any (nonempty) subinterval I_1 of I, the subset of $\mathcal{X}$ of the form $\{\mathbf{x} : \boldsymbol{\eta}'\mathbf{x} \in I_1\}$ has positive probability.

Now let J be an open subinterval of I on which $\mu_w(\cdot)$ is strictly increasing. Let A be the set $\{\mathbf{x} : \boldsymbol{\eta}'\mathbf{x} \in J\}$. Then $\Pr(A) > 0$. Because μ_w is continuous and strictly increasing in J, for any $\mathbf{x}$ in A, the integral

$$\int_{\boldsymbol{\eta}'\mathbf{x}}^{\theta + \mathbf{h}'\mathbf{x}} \{\mu(\boldsymbol{\eta}'\mathbf{x}) - \mu(s)\}\, ds < 0.$$

This is because $\theta + \mathbf{h}'\mathbf{x} \neq \boldsymbol{\eta}'\mathbf{x}$ and there is at least one point t between these two numbers such that $\mu(\boldsymbol{\eta}'\mathbf{x}) - \mu(s) \neq 0$. Hence

$$\Pr\left(\int_{\boldsymbol{\eta}'\mathbf{x}}^{\theta + \mathbf{h}'\mathbf{x}} \{\mu(\boldsymbol{\eta}'\mathbf{x}) - \mu(s)\}\, ds < 0\right) \geq \Pr(A) > 0. \qquad \square$$

PROOF OF PROPOSITION 3.3. The first part follows from display (14). The third part is a consequence of Cook and Li (2002) as applied to individual

subpopulations. We prove the second part. We omit the subscript w in exactly the same way as we did in the proof of Proposition 3.1. First, assume that $\mathrm{E}(\mathbf{X}) = 0$. Let $\boldsymbol{\beta}$ be the population regression coefficient $\boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}Y)$ and let $\mathbf{P} = \mathbf{P}(\boldsymbol{\eta}, \boldsymbol{\Sigma})$ be the projection onto $\mathrm{Span}(\boldsymbol{\beta})$ with respect to the inner product induced by $\boldsymbol{\Sigma}$, that is, $\mathbf{P} = \boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{\Sigma}/(\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})$. It is easy to check that $R(\theta, \mathbf{h})$ can be rewritten in the form

$$R(\theta, \mathbf{h}) = -\theta\mathrm{E}(Y) - \mathbf{h}'\boldsymbol{\Sigma}\boldsymbol{\beta} + \mathrm{E}\phi(\theta + \mathbf{h}'\mathbf{X}). \tag{32}$$

Because $\boldsymbol{\beta} = \mathbf{P}\boldsymbol{\beta}$, the second term above equals

$$-\mathbf{h}'\boldsymbol{\Sigma}\mathbf{P}\boldsymbol{\beta} = -\mathbf{h}'\boldsymbol{\Sigma}\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}Y) = -(\mathbf{P}\mathbf{h})'\mathrm{E}(\mathbf{X}Y),$$

where for the second equality we have used the identity $\boldsymbol{\Sigma}\mathbf{P}\boldsymbol{\Sigma}^{-1} = \mathbf{P}'$. So the first two terms on the right-hand side of (32) can be written as

$$-\mathrm{E}\{Y(\theta + (\mathbf{P}\mathbf{h})'\mathbf{X})\}.$$

By Jensen's inequality the last term on the right-hand side of (32) is no smaller than $\mathrm{E}\phi(\theta + \mathbf{h}'\mathrm{E}(\mathbf{X}|\boldsymbol{\beta}'\mathbf{X}))$, in which $\mathrm{E}(\mathbf{X}|\boldsymbol{\beta}'\mathbf{X})$ equals $\mathbf{P}'\mathbf{X}$ because the former is linear in $\boldsymbol{\beta}'\mathbf{X}$ [see Cook (1998a), page 57]. Hence

$$\mathrm{E}\phi\big(\theta + \mathbf{h}'\mathrm{E}(\mathbf{X}|\boldsymbol{\beta}'\mathbf{X})\big) \geq \mathrm{E}\phi\big(\theta + (\mathbf{P}\mathbf{h})'\mathbf{X}\big).$$

It follows that $R(\theta, \mathbf{h}) \geq R(\theta, \mathbf{P}\mathbf{h})$ for any $(\theta, \mathbf{h})$, which, because $R(\theta, \mathbf{h})$ has a unique minimizer in $\mathbf{h}$, implies that the minimizer is of the form $\mathbf{P}\mathbf{h}$. This gives the desired result because $\mathbf{P}\mathbf{h}$ is parallel to $\boldsymbol{\beta}$. For the case where $\mathrm{E}(\mathbf{X}) \neq 0$, we note that

$$R(\theta, \mathbf{h}) = \mathrm{E}L(\theta + \mathbf{h}'\mathbf{X}, Y) = \mathrm{E}L\big(\theta + \mathbf{h}'\mathrm{E}(\mathbf{X}) + \mathbf{h}'\big(X - \mathrm{E}(\mathbf{X})\big), Y\big).$$

Thus if we let $\theta^* = \theta + \mathbf{h}'\mathrm{E}(\mathbf{X})$ and $\mathbf{X}^* = \mathbf{X} - \mathrm{E}(\mathbf{X})$, then the same argument for the $\mathrm{E}(\mathbf{X}) = 0$ case can be applied to complete the proof. □

For subsequent use in the proof of the large sample results of Section 4, note that the singular value decomposition in (19) implies $\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\mathbf{B}^*\boldsymbol{\Omega}^{-1/2} = 0$ and $\boldsymbol{\Sigma}_{\cdot}^{1/2}\mathbf{B}^*\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Psi} = 0$ or, equivalently,

$$\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\mathbf{B}^* = 0 \quad \text{and} \quad \mathbf{B}^*\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Psi} = 0. \tag{33}$$

Also, note that under (24) we have $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{\cdot}$ for all $w = 1, \ldots, c$ and $\hat{\boldsymbol{\Sigma}}_{\cdot}$ is consistent for $\boldsymbol{\Sigma}_{\cdot}$.

PROOF OF PROPOSITION 4.1. The matrix $\mathbf{U}$ can be expanded as

$$\begin{aligned}\mathbf{U} = {} & \sqrt{n}\boldsymbol{\Gamma}'\big(\hat{\boldsymbol{\Sigma}}_{\cdot}^{1/2} - \boldsymbol{\Sigma}_{\cdot}^{1/2}\big)\mathbf{B}^*\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Psi} + \sqrt{n}\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}(\hat{\mathbf{B}}^* - \mathbf{B}^*)\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Psi} \\ & + \sqrt{n}\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\mathbf{B}^*\big(\hat{\boldsymbol{\Omega}}^{-1/2} - \boldsymbol{\Omega}^{-1/2}\big)\boldsymbol{\Psi} + O_p(n^{-1/2}).\end{aligned}$$

By (33), the first and the last terms are $\mathbf{0}$, so that $\mathbf{U}$ can be approximated by

$$\mathbf{U} = \sqrt{n}\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}(\hat{\mathbf{B}}^* - \mathbf{B}^*)\,\boldsymbol{\Omega}^{-1/2}\,\boldsymbol{\Psi} + O_p(n^{-1/2}).$$

Recall that $\hat{\mathbf{B}}^* - \mathbf{B}^*$ is a matrix with columns $\hat{a}_w\hat{\boldsymbol{\beta}}_w - a_w\boldsymbol{\beta}_w$, $w = 1, \ldots, c$. Since $\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\mathbf{B}^* = 0$, the coefficient a_w in $a_w\boldsymbol{\beta}_w$ can be replaced by an arbitrary number, and in particular by $\hat{a}_w$. This gives

$$\begin{aligned}&\sqrt{n}\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}(\hat{\mathbf{B}}^* - \mathbf{B}^*)\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Psi}\\ &\quad = \boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\big(\sqrt{n_1}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1), \ldots, \sqrt{n_c}(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_c)\big)\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Psi}.\end{aligned}$$

Let us expand $\sqrt{n_w}(\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w)$. It is easy to see that

$$\hat{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\sigma}}_w - \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\sigma}_w = \big(\hat{\boldsymbol{\Sigma}}_w^{-1} - \boldsymbol{\Sigma}_w^{-1}\big)\boldsymbol{\sigma}_w + \boldsymbol{\Sigma}_w^{-1}(\hat{\boldsymbol{\sigma}}_w - \boldsymbol{\sigma}_w) + O_p(n^{-1}). \tag{34}$$

We first expand the difference $\hat{\boldsymbol{\Sigma}}_w^{-1} - \boldsymbol{\Sigma}_w^{-1}$, and for this we need an expansion of $\hat{\boldsymbol{\Sigma}}_w - \boldsymbol{\Sigma}_w$. Note that

$$\begin{aligned}&\sqrt{n_w}(\hat{\boldsymbol{\Sigma}}_w - \boldsymbol{\Sigma}_w)\\ &\quad = n_w^{-1/2}\sum_{i=1}^{n_w}[(\mathbf{X}_{iw} - \bar{\mathbf{X}}_w)(\mathbf{X}_{iw} - \bar{\mathbf{X}}_w)' - \boldsymbol{\Sigma}_w] + O_p(n^{-1/2})\\ &\quad = n_w^{-1/2}\sum_{i=1}^{n_w}\big[\big(\mathbf{X}_{iw} - \mathrm{E}(\mathbf{X}_w)\big)\big(\mathbf{X}_{iw} - \mathrm{E}(\mathbf{X}_w)\big)' - \boldsymbol{\Sigma}_w\big] + O_p(n^{-1/2})\\ &\quad = n_w^{-1/2}\boldsymbol{\Sigma}_w^{1/2}\sum_{i=1}^{n_w}(\mathbf{Z}_{iw}\mathbf{Z}_{iw}' - \mathbf{I})\boldsymbol{\Sigma}_w^{1/2} + O_p(n^{-1/2})\\ &\quad \equiv \boldsymbol{\Delta}_{n,w} + O_p(n^{-1/2}),\end{aligned}$$

where $E(\mathbf{X}_w)$ denotes $E(\mathbf{X}|W = w)$ and $\mathbf{Z}_{iw} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X}_{iw} - \mathrm{E}(\mathbf{X}_w))$ denotes a standardized predictor observation from subsample w. Express the unknown expansion of $\hat{\boldsymbol{\Sigma}}_w^{-1}$ as $\boldsymbol{\Sigma}_w^{-1} + n_w^{-1/2}\mathbf{D}_{n,w} + O_p(n^{-1})$ for some random matrix $\mathbf{D}_{n,w} = O_p(1)$. We let $\mathbf{D}_{n,w}$ be such that $\hat{\boldsymbol{\Sigma}}_w^{-1}\boldsymbol{\Sigma}_w = \mathbf{I} + O_p(n^{-1})$. By simple algebra, this equation is equivalent to

$$\mathbf{D}_{n,w} = -\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Delta}_{n,w}\boldsymbol{\Sigma}_w^{-1} = -n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{i=1}^{n_w}(\mathbf{Z}_{iw}\mathbf{Z}_{iw}' - \mathbf{I})\boldsymbol{\Sigma}_w^{-1/2}.$$

In other words,

$$\hat{\boldsymbol{\Sigma}}_w^{-1} - \boldsymbol{\Sigma}_w^{-1} = -n_w^{-1}\boldsymbol{\Sigma}_w^{-1/2}\sum_{i=1}^{n_w}(\mathbf{Z}_{iw}\mathbf{Z}_{iw}' - \mathbf{I})\boldsymbol{\Sigma}_w^{-1/2} + O_p(n^{-1}). \tag{35}$$

Next, we expand $\hat{\boldsymbol{\sigma}}_w - \boldsymbol{\sigma}_w$. We have

$$
\begin{aligned}
&\sqrt{n_w}(\hat{\boldsymbol{\sigma}}_w - \boldsymbol{\sigma}_w) \\
&\quad = n_w^{-1/2} \sum_{i=1}^{n_w} [(\mathbf{X}_{iw} - \bar{\mathbf{X}}_w)(Y_{iw} - \bar{Y}_w) - \boldsymbol{\sigma}_w] \\
&\quad = n_w^{-1/2} \sum_{i=1}^{n_w} \big[\big(\mathbf{X}_{iw} - \mathrm{E}(\mathbf{X}_w)\big)\big(Y_{iw} - \mathrm{E}(Y_w)\big) - \boldsymbol{\sigma}_w\big] + O_p(n^{-1/2}) \\
&\quad = n_w^{-1/2} \boldsymbol{\Sigma}_w^{1/2} \sum_{i=1}^{n_w} \big[\mathbf{Z}_{iw}\big(Y_{iw} - \mathrm{E}(Y_w)\big) - \boldsymbol{\Sigma}_w^{-1/2} \boldsymbol{\sigma}_w\big] + O_p(n^{-1/2}),
\end{aligned} \tag{36}
$$

where $\mathrm{E}(Y_w)$ denotes the conditional expectation $\mathrm{E}(Y|W = w)$. Now substitute (35) and (36) into (34) to obtain

$$
\begin{aligned}
&\sqrt{n_w}(\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w) \\
&\quad = n_w^{-1/2} \boldsymbol{\Sigma}_w^{-1/2} \sum_{i=1}^{n_w} \big(-\mathbf{Z}_{iw} \mathbf{Z}_{iw}' \boldsymbol{\Sigma}_w^{-1/2} \boldsymbol{\sigma}_w + \mathbf{Z}_{iw}\big(Y_{iw} - \mathrm{E}(Y_w)\big)\big) + O_p(n^{-1/2}) \\
&\quad = n_w^{-1/2} \boldsymbol{\Sigma}_w^{-1/2} \sum_{i=1}^{n_w} \mathbf{Z}_{iw} e_{iw} + O_p(n^{-1/2}),
\end{aligned}
$$

where $e_{iw} = (Y_{iw} - \mathrm{E}(Y_w)) - \boldsymbol{\sigma}_w' \boldsymbol{\Sigma}_w^{-1/2} \mathbf{Z}_{iw}$. Note that, by the definitions of $\boldsymbol{\sigma}_w$ and $\mathbf{Z}_{iw}$, the quantity e_{iw} is in fact the residual for the linear regression of Y_{iw} on $\mathbf{X}_{iw}$ within the subpopulation w, as we defined in (17). It follows that

$$
\mathbf{U} = \sum_{w=1}^{c} \boldsymbol{\Gamma}' \boldsymbol{\Sigma}_{\cdot}^{1/2} \boldsymbol{\Sigma}_w^{-1/2} \left(n_w^{-1/2} \sum_{i=1}^{n_w} \mathbf{Z}_{iw} e_{iw} \right) \omega_w^{-1/2} \boldsymbol{\psi}_w'. \tag{37}
$$

Taking the vec of $\mathbf{U}$, we thus have

$$
\mathrm{vec}(\mathbf{U}) = \sum_{w=1}^{c} \big(\omega_w^{-1/2} \boldsymbol{\psi}_w \otimes \boldsymbol{\Gamma}' \boldsymbol{\Sigma}_{\cdot}^{1/2} \boldsymbol{\Sigma}_w^{-1/2}\big) \left(n_w^{-1/2} \sum_{i=1}^{n_w} \mathbf{Z}_{iw} e_{iw} \right).
$$

An application of the central limit theorem proves the desired result. □

PROOF OF PROPOSITION 4.2. Let $\mathbf{P}_w = \boldsymbol{\Sigma}_w^{1/2} \boldsymbol{\beta}_w (\boldsymbol{\beta}_w' \boldsymbol{\Sigma}_w \boldsymbol{\beta}_w)^{-1} \boldsymbol{\beta}_w' \boldsymbol{\Sigma}_w^{1/2}$ be the projection matrix onto the space spanned by $\boldsymbol{\Sigma}_w^{1/2} \boldsymbol{\beta}_w$ and let $\mathbf{Q}_w = \mathbf{I} - \mathbf{P}_w$. By (33), the columns of $\boldsymbol{\Gamma}' \boldsymbol{\Sigma}_{\cdot}^{1/2}$ are orthogonal to the vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_c$. Hence $\boldsymbol{\Gamma}' \boldsymbol{\Sigma}_{\cdot}^{1/2} \boldsymbol{\Sigma}_w^{-1/2} \mathbf{P}_w = 0$ or, equivalently, $\boldsymbol{\Gamma}' \boldsymbol{\Sigma}_{\cdot}^{1/2} \boldsymbol{\Sigma}_w^{-1/2} \mathbf{Q}_w = \boldsymbol{\Gamma}' \boldsymbol{\Sigma}_{\cdot}^{1/2} \boldsymbol{\Sigma}_w^{-1/2}$. So we can insert a $\mathbf{Q}_w$ in front of the $\mathbf{Z}_{iw}$ in (37) without changing the quantity, that is,

$$
\mathbf{U} = \sum_{w=1}^{c} \boldsymbol{\Gamma}' \boldsymbol{\Sigma}_{\cdot}^{1/2} \boldsymbol{\Sigma}_w^{-1/2} \left(n_w^{-1/2} \sum_{i=1}^{n_w} \mathbf{Q}_w \mathbf{Z}_{iw} e_{iw} \right) \omega_w^{-1/2} \boldsymbol{\psi}_w'.
$$

It follows that vec(**U**) has an asymptotic normal distribution with mean **0** and covariance matrix

$$\begin{aligned}(38)\qquad &\sum_{w=1}^{c}\bigl(\omega_w^{-1/2}\boldsymbol{\psi}_w\otimes\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Sigma}_w^{-1/2}\bigr)\\ &\quad\times \mathrm{E}(e_w^2\mathbf{Q}_w\mathbf{Z}_w\mathbf{Z}_w'\mathbf{Q}_w)\bigl(\omega_e^{-1/2}\boldsymbol{\psi}_w'\otimes\boldsymbol{\Sigma}_w^{-1/2}\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Gamma}\bigr).\end{aligned}$$

Now, since $\mathcal{S}_{Y_w|\mathbf{X}_w}=\mathrm{Span}(\boldsymbol{\beta}_w)$ by assumption, we have $\mathrm{E}(e_w^2|\mathbf{Z}_w)=\mathrm{E}(e_w^2|\mathbf{P}_w\mathbf{Z}_w)$. Therefore, the expectation in the middle of (38) becomes

$$\begin{aligned}(39)\qquad \mathrm{E}(e_w^2\mathbf{Q}_w\mathbf{Z}_w\mathbf{Z}_w'\mathbf{Q}_w)&=\mathrm{E}\bigl(\mathrm{E}(e_w^2|\mathbf{P}_w\mathbf{Z}_w)\mathbf{Q}_w\mathbf{Z}_w\mathbf{Z}_w'\mathbf{Q}_w\bigr)\\ &=\mathrm{E}\bigl(e_w^2\mathrm{E}(\mathbf{Q}_w\mathbf{Z}_w\mathbf{Z}_w'\mathbf{Q}_w|\mathbf{P}_w\mathbf{Z}_w)\bigr)\\ &=\omega_w\mathbf{Q}_w=\omega_w\otimes\mathbf{Q}_w,\end{aligned}$$

where the third equality follows from the normality assumption on $\mathbf{X}_w$ and thus $\mathbf{Z}_w$, and the orthogonality between $\mathbf{P}_w$ and $\mathbf{Q}_w$. As a consequence, the matrix in (38) reduces to

$$\begin{aligned}&\sum_{w=1}^{c}\bigl(\omega_w^{-1/2}\boldsymbol{\psi}_w\otimes\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Sigma}_w^{-1/2}\bigr)(\omega_w\otimes\mathbf{Q}_w)\bigl(\omega_w^{-1/2}\boldsymbol{\psi}_w'\otimes\boldsymbol{\Sigma}_w^{-1/2}\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Gamma}\bigr)\\ &\quad=\sum_{w=1}^{c}\bigl(\boldsymbol{\psi}_w\boldsymbol{\psi}_w'\bigr)\otimes\bigl(\boldsymbol{\Gamma}'\,\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{\cdot}^{1/2}\boldsymbol{\Gamma}\bigr)\end{aligned}$$

as desired. □

Acknowledgments. We are grateful to two referees and an Associate Editor for their prompt and insightful reviews that helped to improve the manuscript significantly.

REFERENCES

BENTLER, P. M. and XIE, J. (2000). Corrections to test statistics in principal Hessian directions. *Statist. Probab. Lett.* **47** 381–389.

BURA, E. and COOK, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 393–410.

CHIAROMONTE, F. and COOK, R. D. (2002). Sufficient dimension reduction and graphics in regression. *Ann. Inst. Statist. Math.* **54** 768–795.

CHIAROMONTE, F., COOK, R. D. and LI, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.* **30** 475–497.

COOK, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91** 983–992.

COOK, R. D. (1998a). *Regression Graphics*. Wiley, New York.

COOK, R. D. (1998b). Principal Hessian directions revisited. *J. Amer. Statist. Assoc.* **93** 84–100.

COOK, R. D. and LEE, H. (1999). Dimension reduction in binary response regression. *J. Amer. Statist. Assoc.* **94** 1187–1200.

COOK, R. D. and LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** 455–474.

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction." *J. Amer. Statist. Assoc.* **86** 28–33.

COOK, R. D. and WEISBERG, S. (1999). *Applied Regression Including Computing and Graphics.* Wiley, New York.

EATON, M. L. and TYLER, D. E. (1994). The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis. *J. Multivariate Anal.* **50** 238–264.

FOULADI, R. T. (1997). Type I error control of some covariance structure analysis technique under conditions of multivariate non-normality. *Comput. Statist. Data Anal.* **29** 526–532.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342.

LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039.

LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052.

SATTERTHWAITE, F. E. (1941). Synthesis of variance. *Psychometrika* **6** 309–316.

SMITH, J. W., EVERHART, J. E., DICKSON, W. C., KNOWLER, W. C. and JOHANNES, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proc. Twelfth Annual Symposium on Computer Applications in Medical Care* 261–265. IEEE Computer Society Press, New York.

B. LI
F. CHIAROMONTE
DEPARTMENT OF STATISTICS
THE PENNSYLVANIA STATE UNIVERSITY
326 THOMAS BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
E-MAIL: bing@stat.psu.edu

R. D. COOK
SCHOOL OF STATISTICS
1994 BUFORD AVE.
UNIVERSITY OF MINNESOTA
ST. PAUL, MINNESOTA 55108